Prognostic Factors in Spinal Cord Injury Clinical Trials

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Lorenzo G. Tanadini

von

Losone TI

Promotionskommission

Prof. Dr. Torsten Hothorn (Vorsitz)
Prof. Dr. med. Armin Curt
Prof. Dr. Leonhard Held

Zürich, 2017

Rerum cognoscere causas

— Virgil

"To Know the Causes of Things"
From verse 490 of Book 2 of the Georgics, composed in 29 BC by the Latin poet Virgil.

Preface

This dissertation has been developed within the Structured Ph.D. Program "Epidemiology and Biostatistics" of the University of Zurich between April 2013 and October 2016. Its fundament resulted from a collaborative effort between the Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute, and the Spinal Cord Injury Center at Balgrist University Hospital. A year spent as a visiting research student at the Department of Statistics of the London School of Economics provided the opportunity to further develop its content. Its core scientific output are the four articles bundled in this thesis.

This cooperative project and my continuing scientific and statistical education could not have taken place without the involvement of several people who I would like to thank for their precious support and guidance throughout this endeavour. Firstly, I would like to thank my main supervisor Torsten Hothorn for his invaluable help during the entire project. His support was always characterised by a profound matter understanding. Many thanks also to my clinical supervisor Armin Curt, who triggered the steps leading to the start of my dissertation, for his continuous assistance and contributions regarding clinical aspects of my research. In this sense, I would like to gratefully mention John Steeves (ICORD, University of British Columbia and Vancouver Coastal Health), who took over a considerable supervision role during the initial stage of my dissertation. I would also like to thank Irini Moustaki for her active supervision and valuable discussions during my visiting period at the Department of Statistics of the London School of Economics. Further, I would like to thank Leonhard Held, who acted as a committee member and provided helpful inputs and suggestions during our meetings, and Prof. Olaf Gefeller (Friedrich-Alexander-University Erlangen-Nürnberg) for acting as external reviewer and providing insightful comments on the submitted thesis.

I would also like to thank the European Multicenter Study about Spinal Cord Injury for granting access to their database tracking the neurological recovery of thousands of patients. My thesis capitalised on their industrious work. I am also grateful to three foundations for their financial support: the International Foundation for Research in Paraplegia provided means to cover two years of salary, the Swiss National Science Foundation granted a doctoral mobility fellowship to visit Prof. Moustaki at LSE in London, UK, and the Jangge-Phön Foundation made important contributions towards the concluding phase of my thesis.

Many thanks also go to all friends, colleagues and staff members of the Biostatistics Department and the Spinal Cord Injury Center. Too many to be listed by name, they nonetheless all contributed to a stimulating research environment and an enjoyable working atmosphere.

Most importantly, I would like to thank my wife Jasmin, little Levin and my family for much appreciated support through good and less good phases of my dissertation.

Frick, October 2016

Lorenzo G. Tanadini

Zusammenfassung

Die rezente Abwanderung der Pharmaunternehmen aus der neurologischen Forschung als Folge einer grossen Zahl gescheiterter klinischer Studien mit Tausenden von Patienten hat die Entwicklung neuer therapeutischer Ansätze stark gebremst. Auf dem Gebiet der Rückenmarksverletzung ist die Situation noch gravierender: Patienten haben bisher weder einen Zugang zu spezifischen Behandlungen, die die Verbesserung ihrer neurologischen Funktionen zum Ziel haben, noch gibt es einen weltweit akzeptierten Versorgungsstandard.

Diese Verögerung ist zum Teil darauf zurückzuführen, dass klinische Studien seltener neurologischer Störungen zu den schwierigsten durchzuführenden Studien überhaupt zählen. Aufbauend auf den Meilensteinen historischer Arbeiten, befasst sich die vorliegende Dissertation mit zwei Hauptmerkmalen klinischer Studien von Rückenmarksverletzungen, nämlich der hohen Variabilität der Erholungsmuster bei geringer Anzahl potenzieller Teilnehmer, sowie der Analyse von facettenreichen neurologischen Endpunkten. Beide Aspekte gelten schon lange als entscheidend für die Entwicklung effizienterer und wirksamerer Studien von Rückenmarksverletzungen und verwandten Disziplinen. In diesem Zusammenhang bedarf es im Speziellen der Annahme innovativer und strengerer statistischer Methoden, die die derzeit eingesetzten, aber fehlerhaften Ansätze ersetzen sollen.

Wir suchten zunächst einen Ansatz für die zuverlässige Identifizierung und Stratifizierung homogener Untergruppen innerhalb einer heterogenen Patientenpopulation. Durch die erstmalige Anwendung einer rekursiven Partitionierungstechnik, die auf bedingter Inferenz basiert, wurde eine innovative Lösung entwickelt, die auf einer eigens dafür entwickelten statistischen Methode basiert. Die Partitionierungstechnik soll die Planung künftiger klinischer Versuche durch die Bereitstellung datengetriebener Entscheidungsregeln für Patienteneinschluss / -ausschluss und -stratifizierung verbessern. Ein Vergleich mit gängigen Stratifikationsmethoden hebt ausserdem mehrere Vorteile der rekursiven Partitionierung hervor. In einem weiteren Schritt haben wir eine umfassende Analyse und Validierung unseres Ansatzes implementiert, die eine flexible Vorlage für zukünftige Studien bietet.

In einem abschliessenden Teil der Arbeit behandelten wir die Analyse komplexer kategorischer Endpunkte, die allgegenwärtig in der neurologischen Forschung und Rückenmarksverletzungsstudien sind. Während Routineanalyseansätze annehmen, dass sogenannte Multiple-Item-Rating-Skalen kontinuierliche Endpunkte liefern, sind diese Ansätze durch starke Einschränkungen fehlerhaft. Wir haben zunächst ein ordinales Regressionsmodell auf die spezifischen Bedürfnisse von klinischen Studien von Rückenmarksverletzungen ausgerichtet, die zu erheblichen Verbesserungen in Bezug auf die statistische Aussagekraft und klinische Interpretation führten. Schliesslich beurteilten wir die Messeigenschaften von Rating-Skalen und deren Langzeitmodellierung auf der Basis von latenten Variablenmodellen, die ein wesentlich strengeres und anspruchsvolleres Rahmenwerk für die Analyse neurologischer Endpunkte liefern.

Durch die Erarbeitung wichtiger Aspekte der Patientenintegration und deren Stratifizierung, sowie der Endpunktanalyse zielten wir darauf ab, Planung und Analyse klinischer Studien zu verbessern, insbesondere die der Rückenmarksverletzungen, sowie neurologischer Störungen im Allgemeinen. Unsere Hoffnung ist, dass die neu erarbeiteten Ansätze synergistisch den klinischen Prozess schnell voranbringen können und so zu spezifischen therapeutischen Behandlungen für Menschen mit Rückenmarksverletzungen führen werden.

Abstract

The departure of drug companies from neurological research as a consequence of a large number of failed clinical trials involving thousands of patients strongly braked the development of new therapeutic approaches. The situation is even more acute in the field of spinal cord injury, where patients have yet to gain access to a first specific treatment for improving neurological function or witness the introduction of a globally accepted standard of care.

The lagging translation process is due to the fact that clinical trials in rare neurological disorders are among the most challenging to conduct. Building on the important lessons provided by historical trials in the path to translation, this thesis addresses two major peculiarities of spinal cord injury trials, namely the high variability of recovery patterns among the small number of potential participants and the analysis of multi-faceted neurological endpoints. Both aspects have been recognised as crucial for the development of more efficient and effective trials in spinal cord injury and related disciplines. In this context, the adoption of innovative and more rigorous statistical methods replacing currently employed but flawed approaches is required.

We first seek an approach for the reliable identification and stratification of homogeneous subgroups within a heterogeneous patient population. By applying for the first time in the field an unbiased recursive partitioning technique based on conditional inference, we provided an innovative solution based on a statistical method specifically developed for this purpose. This partitioning technique is intended to improve the planning of future clinical trials by providing data-driven decision rules for patient inclusion/exclusion and stratification. A comparison with commonly employed stratification methods highlights several advantages of recursive partitioning. In a further step, we implemented a comprehensive analysis and validation of our approach, providing a flexible template applicable to future studies.

Subsequently, we dealt with the analysis of complex categorical endpoints, which are ubiquitous in neurology and spinal cord injury. While routine analysis approaches assume multiple-item rating scales to deliver continuous endpoints, these approaches are flawed by severe limitations. We first adapted an ordinal regression model to the specific needs of spinal cord injury trials, which lead to substantial improvements in terms of statistical power and clinical interpretation. Finally, we assessed measurement properties of rating scales and their longitudinal modelling based on latent variable models, providing a much more rigorous and sophisticated inferential framework for the analysis of neurological outcomes.

By addressing both key issues of patient inclusion and stratification as well as endpoint analysis, we aimed at providing essential elements for a better planning and analysis of clinical trials in spinal cord injury in particular, and neurological disorders in general. Our hope is that they may synergistically fast-forward the clinical process leading to specific therapeutic treatments for people living with spinal cord injury.

Thesis outline

Introduction

Paper I: **Identifying Homogeneous Subgroups in Neurological Disorders: Unbiased Recursive Partitioning in Cervical Complete Spinal Cord Injury**

Lorenzo G. Tanadini, John D. Steeves, Torsten Hothorn, Rainer Abel, Doris Maier, Martin Schubert, Norbert Weidner, Rüdiger Rupp, Armin Curt

Published in *Neurorehabilitation and Neural Repair*, 2014, **Vol. 28** (6), 507-515.

Paper II: **Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders: Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury**

Lorenzo G. Tanadini, Torsten Hothorn, Linda A. T. Jones, Daniel P. Lammertse, Rainer Abel, Doris Maier, Rüdiger Rupp, Norbert Weidner, Armin Curt, John D. Steeves

Published in *Neurorehabilitation and Neural Repair*, 2015, **Vol. 29** (9), 867-877.

Paper III: **Autoregressive transitional ordinal model to test for treatment effect in neurological trials with complex endpoints**

Lorenzo G. Tanadini, John D. Steeves, Armin Curt, Torsten Hothorn

Published in *BMC Medical Research Methodology*, 2016, 16-149.

Paper IV: **Addressing limitations of rating scales and their analysis in spinal cord injury under the unifying framework of latent variable modeling**

Lorenzo G. Tanadini, Armin Curt, Irini Moustaki

Submitted manuscript.

Introduction

Just as the scientific understanding of brain plasticity has reached new heights, drug companies started turning their backs to neurological disorders. A number of failed clinical trials involving thousands of patients prompted companies to divert their investments to other medical fields. While the failure of any clinical trial can be usually ascribed to multiple reasons, the poor choice of primary endpoints and their analysis pose a common threat to neurological studies. In addition, rare neurological diseases such as spinal cord injuries are further characterised by few potential patients, but a large heterogeneity in terms of spontaneous recovery patterns. Clinical trials in rare neurological diseases are therefore among the most challenging to conduct. The complexity of the situation is reflected by the fact that after three decades of clinical research, there is still not an approved treatment for improving neurological function nor a consensus standard of care specific to people who suffered a spinal cord injury.

A historical review of mainly failed spinal cord trials in the path to translation has nonetheless provided important lessons for improving future trials conduct. Especially trial design, the choice of the primary outcome and its analysis plan were identified as key elements that require the application of innovative and more rigorous statistical approaches. This thesis pursued thus two aims which are meant to synergistically contribute to the development of more efficient and effective clinical trials in spinal cord injury and related disciplines.

Our first aim was to propose an innovative approach for the reliable stratification of homogeneous participant subgroups within a heterogeneous neurological disorder such as spinal cord injury. Given the additional constraints posed by the rare disease status, a necessary condition was to propose a framework which allowed to include as many participants as possible in a sensible manner. This was achieved in Paper I and Paper II of this thesis, where we applied for the first time in the field the unbiased recursive partitioning technique called conditional inference tree to improve the planning of future clinical trials by providing data-driven decision rules for patient stratification and inclusion/exclusion.

Our second aim was to further develop and tailor statistical methods used to analyse complex ordinal endpoints generated by multiple-item rating scales. Although this kind of endpoint is ubiquitous in neurology and spinal cord injury, commonly used analysis approaches routinely assume more refined measurement scales (e.g. continuous endpoints). These approaches introduce several limitations and represent a major weakness of the translational process. To address this issue, Paper III and Paper IV propose two ordinal models that represent specific solutions to the analysis of complex ordinal endpoints and highlighted drawbacks of currently employed approaches in spinal cord injury.

All analyses conducted in this thesis are intended to promote more efficient and effective clinical trials by overcoming major weaknesses of the current translational process in spinal cord injury. Both key areas of patient stratification/inclusion as well as endpoint analysis were addressed with innovative and more rigorous statistical approaches, which may further act as templates for similar scientific endeavours across medical disciplines.

1 Neurological research

Neurological research is responsible for the investigation of many devastating disorders such as stroke, Alzheimer and Parkinson, among others. Besides impairment and sufferance, brain-related disorders are a greater socio-economic burden than cancer, cardiovascular diseases and diabetes combined (Gustavsson *et al.*, 2011). With yearly costs for the European society estimated at almost 400 billion € (Andlin-Sobocki *et al.*, 2005), the economic potential of any drug obtaining approval from regulatory agencies set a strong incentive for drug development. In fact, the translational process, which starts from promising experiments in the laboratory and results in effective treatments safely applicable to humans, experienced a sustained acceleration at the turn of the century.

More recently, though, health insurers and drug companies have withdrawn from neuroscience as a consequence of a large number of failed trials involving thousands of patients (Schwab and Buchli, 2012). This unfortunate development occurred despite the concomitant unparalleled evolution of preclinical discoveries with high potential for translation to human cure. Mechanisms of cellular and molecular regeneration occurring at the lesion site, and the possibility to actively regulate them, have created the premises for several therapeutic approaches (Thuret *et al.*, 2006; Tator, 2006; Hawryluk *et al.*, 2008; Liu *et al.*, 2011). Function-blocking antibodies (Zörner and Schwab, 2010), stem cell therapies (Antonic *et al.*, 2013), and growth-promoting immunotheraphy (Wahl *et al.*, 2014) are some of the most recent investigatory lines potentially leading to human therapies.

Given all these research lines generating very promising outcomes in more controlled settings, the question arises as why those approaches have failed to deliver treatments in the clinical context. It is certainly true that the failure of any clinical study cannot be reduced to a single aspect. Nonetheless, the choice of a primary endpoint and its analysis plan are very often criticised for being crude, not well defined, or poorly analysed (Hobart, 2003; Bath *et al.*, 2007; Hobart *et al.*, 2007; Lammertse, 2012; Schwab and Buchli, 2012; Maas *et al.*, 2013). We therefore maintain that a prominent weakness of the translational process is to be located in the statistical analysis methods used to test for treatment efficacy, especially concerning multiple-item rating scales used to measure health outcomes in patients with neurological disorders.

1.1 Multiple-item rating scales are ubiquitous

Virtually all routinely performed clinical assessments in neurological diseases and spinal cord injury are multiple-item questionnaires. Consequently, virtually all potential primary and secondary endpoints are delivered by multiple-item rating scales as well. Above and beyond considerations related to the clinical validity of this type of endpoints, this data format is characterized by an arbitrary numerical scale merely establishing a ranking of observations at item level. This introduces often overlooked statistical properties that, if not considered properly, have the potential to invalidate analyses.

Firstly, despite the numerical labels in the form of successive integers assigned to them, the difference between two following ranks is by no means bound to be equivalent across the range of the rating scale (Hobart, 2003). This feature prevents standard mathematical operations such as addition and multiplication, and makes the use of statistical methods developed for continuous outcomes inappropriate already at item level. Secondly, multiple-item rating scales are hardly ever unidimensional measurement tools (Hobart *et al.*, 2007). Instead, neurological multiple-item rating scales usually track a combination of several health domains.

Even Patient Reported Outcomes are usually analysed as a measure of a single health domain, although they have specifically been developed to capture patients' multifaceted perspective (Chow *et al.*, 2009). While this may seem to be a rather theoretical issue with little relevance for clinical practice, it is akin to having a "scale that measures length at one end, weight in the middle, and volume at the other end" (Hobart, 2003). Especially with regard to the issues outlined above, assuming continuous interval-scaled endpoints in the presence of rating scales has been shown to be inappropriate in a number of aspects (Agresti, 2010). Biased parameter estimates, misleading associations and loss of power are some of the known consequences of its disregard (Winship and Mare, 1984; Hastie *et al.*, 1989; Scott *et al.*, 1997).

1.2 An overall score is commonly adopted, but generally not valid

Nonetheless, measurements from multiple-item rating scales are usually taken at face values and combined into a summed total score. The latter is then handled as a single, interval-scaled endpoint for analysis (Bracken *et al.*, 1984, 1990; Geisler *et al.*, 1991; Bracken *et al.*, 1997; Cardenas *et al.*, 2007). Despite being heavily employed, the practice of adding several ordinal items to form a single overall score is generally not valid with regard to the assumptions of unit change and unidimensionality exemplified above, and has been repeatedly reported in neurological and related physical functioning settings (McHorney *et al.*, 1997; Ravaud *et al.*, 1999; Fink *et al.*, 1999; Luther *et al.*, 2006; Catz *et al.*, 2007; Hobart *et al.*, 2007). The issue of inappropriate statistical analyses of ordinal data delivered by rating scales is not new in medicine (Forrest and Andersen, 1986), but has assumed an unprecedented magnitude in neurology, where ordinal measurements generated by multiple-item rating scales are often the only type of data which clinical assessments provide and therefore what neurologists have to work with (Hobart, 2003). Yet, the rapid spread of questionable analysis approaches cannot be justified on scientific terms, as the statistical framework for the analysis of complex ordinal endpoints is already in place (Agresti, 2010; Tutz, 2012), and has been rapidly expanding ever since the formulation of the first approach to ordinal regression models (the proportional odds model by McCullagh, 1980). Especially the social sciences have been very active in this area of statistics, as categorical and particularly ordinal outcomes are the hallmark of their field (De Boeck and Wilson, 2004; von Davier and Carstensen, 2007). Similar approaches have been sporadically proposed in the field of medical statistics (Laffont *et al.*, 2014; Conigliani *et al.*, 2014), but are rarely applied in neurological and clinical research.

2 Spinal cord injury

In addition to the issues of crude endpoints and flawed analysis strategies common to all neurological disciplines, rare neurological diseases such as spinal cord injury are further characterised by few patients that can be potentially recruited for a study (Wyndaele and Wyndaele, 2006), but a large heterogeneity in terms of spontaneous recovery patterns (Fawcett *et al.*, 2006). This peculiarity of spinal cord studies may lead to distortions of trial results, as the contribution of a small number of participants with heterogeneous characteristics may drive efficacy testing and outcome interpretation in one direction, independently of the real value of the treatment. Also, subtle treatment effects could be overcast by the inherent heterogeneity of recovery patterns, leading to premature abandonment of promising research lines. To prevent these deleterious issues, but also to accommodate the limited number of potentially appropriate trial participants available for an ever increasing number of new interventions, the recruitment of participants needs to be developed in a more inclusive and effective manner.

2.1 Inclusion/exclusion and stratification of trial participants

While applying narrow inclusion criteria may be justified in early trial phases focusing on safety, their enforcement implies slow trial progress, with negative consequences for the timely and financially sustainable accomplishment of the planned goals. In addition, a narrow recruitment strategy is inefficient, as it provides results only applicable to a narrow subpopulation, requiring therefore additional studies for spinal cord participants with other characteristics. Therefore, once treatment safety is established, a more inclusive strategy that enrols participants with varying degrees and sites of lesions should be preferred. Such a strategy is more efficient and generates knowledge that concerns a large patient population. In addition, participants with incomplete lesions are more likely to benefit from interventions targeting their spared sensorimotor function (Tuszynski *et al.*, 2006). The inclusion of participants with incomplete lesions is further justified by the fact that many experimental treatments being translated were developed using animal models with incomplete lesions (Kwon *et al.*, 2002). There are therefore a number of important reasons to develop an inclusive enrolment strategy that will allocate the few trial participants available efficiently, ensuring that a large scientific gain is brought about by the timely and financially sustainable completion of clinical studies.

However, the enrolment of participants with varying degrees of lesions poses specific challenges (Tuszynski *et al.*, 2006). Besides safety challenges, participants with incomplete spinal cord lesions comprise a highly heterogeneous population in terms of level and severity of injury, as well as the diversity of recovery patterns (Fawcett *et al.*, 2006). Historical trials with broad inclusion criteria (Bracken *et al.*, 1984, 1990; Geisler *et al.*, 1991; Bracken *et al.*, 1997; Geisler *et al.*, 2001b; Dobkin *et al.*, 2006; Cardenas *et al.*, 2007) emphasised the necessity to implement stratification algorithms to limit subject heterogeneity within study cohorts. In addition, the exclusion of those patients whose spontaneous recovery is so extensive that would mask any therapeutic benefit must be guaranteed. In summary, the ability to capitalise on a more inclusive recruitment approach requires the ability by trial scientists to prospectively single out clinically relevant subgroups of patients. This capability not only would allow the targeted inclusion/exclusion of participants and provides an a-priori stratification template, but also enables the formulation of strata-specific endpoints in future trials.

3 Fast-forward the lagging translational process

Despite the optimism sparked by the understanding of mechanisms of neuronal degeneration and regeneration (Kleitman, 2004), the promises of preclinical discoveries have yet to be translated into a standard care of treatment (Lammertse, 2012). To consolidate the efforts of the field to continually improve the conception of clinical trials, the International Campaign for Cures of Spinal Cord Injury Paralysis (campaignforcure.org) appointed in 2007 an international panel with the task of reviewing strengths and weaknesses of past clinical trials in spinal cord injury. Their recommendations were condensed in four publications, addressing issues such as patients' spontaneous recovery patterns (Fawcett *et al.*, 2006), clinical trial endpoints (Steeves *et al.*, 2006), trial inclusion and exclusion criteria (Tuszynski *et al.*, 2006), and clinical trial design (Lammertse *et al.*, 2006). While providing important guidelines that effectively influenced the conception of following clinical trials (Sorani *et al.*, 2012), the reviews did not solicit the application of the most appropriate and rigorous statistical approaches with regard to patient stratification and trial outcome analysis. This thesis tackles both issues of patient inclusion as well as endpoint analysis, providing therefore not only much-needed principles for decision-making in the clinical setting, but also essential, yet missing elements for a bet-

ter planning and analysing of clinical trials in neurological disorders. Ultimately, our results are intended to fast-forward the clinical process leading to specific therapeutic treatments for people living with spinal cord injury.

4 Upper Extremity Motor Scores as primary endpoint

In our specific neurological setting, the trial endpoint considered is the Upper Extremity Motor Score (UEMS). UEMS represents a subset of the International Standards for Neurological Classification of Spinal Cord Injury and describes the muscle contraction force for 10 key muscles on the arms and hands (see Figure 1). Each key muscle is rated on a 6-point ordinal scale (0: total paralysis, through 5: active movement against full resistance) (Kirshblum *et al.*, 2011) . Accordingly, $Y_{i,m,t}$ is the muscle contraction score for patient i ($i = 1, ..., n$) and key muscle m ($m = 1, ..., 10$) measured at time point t ($t = 1, ..., 5$). Each key muscle $Y_{i,m,t}$ is therefore an indicator with $k = 6$ levels $0 < 1 < ... < 5$, and UEMS is a multiple-item rating scale. The chosen endpoint is particularly relevant in spinal cord injury research. Total UEMS or its change over trial period has been employed repeatedly in clinical trials (Bracken *et al.*, 1984, 1990; Geisler *et al.*, 1991; Bracken *et al.*, 1997; Cardenas *et al.*, 2007; Casha *et al.*, 2012) and has been suggested to correlate with changes in activities of daily living that rely on recovery of upper extremity function (Rudhe and van Hedel, 2009).

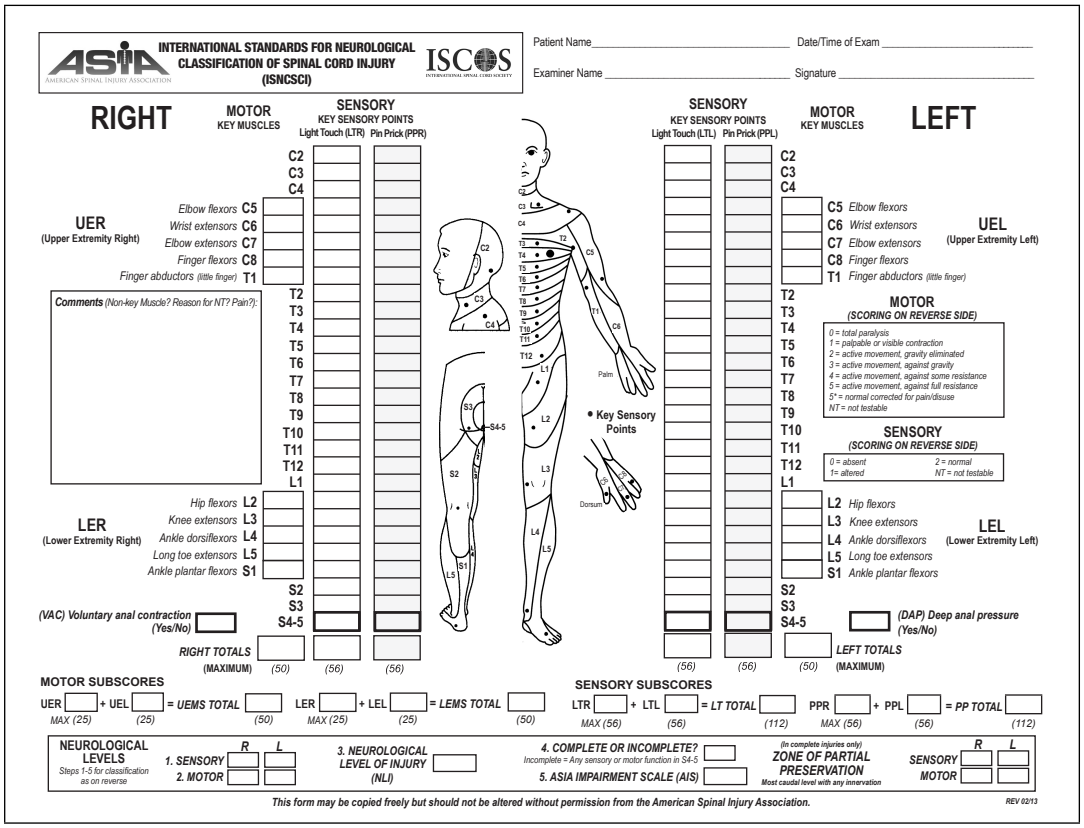

Figure 1.: International Standards for Neurological Classification of Spinal Cord Injury

All the analyses conducted in this thesis are based on data kindly provided by The European Multicenter study about Spinal Cord Injury (emsci.org, ClinicalTrials.gov). The European Multicenter study about Spinal Cord Injury encompasses 19 paraplegic centres across Europe

with the aim to foster close collaboration to discuss, plan and realise prospective studies involving participants who suffered a spinal cord lesion. Patients deferred to member centres who comply with clearly defined inclusion criteria are tested and documented within a fixed time schedule (1, 4, 12, 24 and 48 weeks) after spinal cord injury. The European Multicenter study about Spinal Cord Injury has been tracking the functional, neurological and neurophysiological status of patients during the first year of recovery in a rigorously standardized manner since 2001. The papers composing this thesis are therefore retrospective analyses of prospectively collected neurological data.

5 Planning a study: inclusion/exclusion and stratification

Given the limited pool of individuals who suffer spinal cord injury (Sekhon and Fehlings, 2001), but also the necessity to include participants with varying degrees of injuries to match preclinical studies (Kwon *et al.*, 2002) and maximise the scientific gain for each completed trial, the need to design more inclusive trials is particularly pressing (Tuszynski *et al.*, 2006). However, this can only be sensibly implemented with innovative stratification approaches that predictively identify homogeneous subgroups of trial participants within a heterogeneous population and possibly help define reasonable cohort-specific outcomes.

5.1 Current approaches

A historical revision of key studies in spinal cord injury (Lammertse, 2012) reveals that stratification procedures have been either very narrow (Lammertse *et al.*, 2012), or rather broad (Geisler *et al.*, 2001b; Dobkin *et al.*, 2006; Cardenas *et al.*, 2007), when not basically omitted (Bracken *et al.*, 1984, 1990, 1997; Geisler *et al.*, 1991). Trial scientists have been heavily relying on clinical scales for stratifications (Geisler *et al.*, 2001b; Dobkin *et al.*, 2006; Cardenas *et al.*, 2007), the most common of which is the American Spinal Injury Association Impairment Scale (Kirshblum *et al.*, 2011). The 5-step scale represents a clinical tool developed to broadly categorize motor and sensory impairment in individuals who suffered a spinal cord lesion. Despite their widespread use in the clinical setting, the grades of the Impairment Scale have been repeatedly proven neither to be a valuable measure for endpoint definition (Geisler *et al.*, 2001b; Lammertse *et al.*, 2012; Steeves *et al.*, 2006) nor to suffice as a fine-grained stratification as required by clinical trials (Velstra *et al.*, 2014; Tanadini *et al.*, 2015).

More recently, other researchers attempted to create clinical algorithms for the prediction of long-term outcomes and for patient stratification (Zörner *et al.*, 2010; van Middendorp *et al.*, 2011; Wilson *et al.*, 2012). Nonetheless, their approaches were developed with regard to very specific functional outcomes. In addition, by relying on the statistical modelling techniques of multiple linear and logistic regressions, they left the identification of more homogeneous subgroups for participants with varying degrees of lesions unanswered.

5.2 Unbiased recursive partitioning by conditional inference

To overcome previous limitations and address both the need for inclusive trials but also the necessity of stratification of a heterogeneous spinal cord patient population, we adopted for the first time in the field the approach of recursive partitioning by conditional inference (Hothorn *et al.*, 2006). Tree-based regression models of this type are in fact specifically designed and particularly useful for screening heterogeneous populations to identify more homogeneous patient subgroups.

Generally speaking, the recursive partitioning algorithm is a tree-structured regression model based on sequential tests of independence, sequentially producing binary splits into the initially heterogeneous population to produce disjoint and more and more homogeneous pairs of subgroups (see Figure 2).

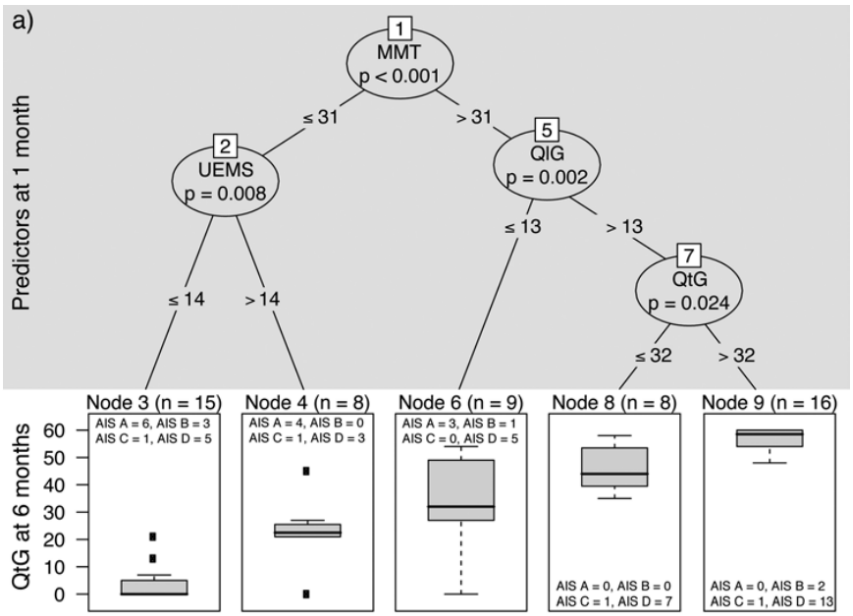

Figure 2.: A clinical application of unbiased recursive partitioning by conditional inference (Figure 1A from Velstra *et al.* (2014)). In an attempt to predict quantitative grasping [QtG] six months after spinal cord injury, the algorithm led to a partition of the initial patient population into five patient subgroups with increasingly positive outcomes. The splitting procedure is based on the selected covariates manual muscle testing [MMT], upper extremity motor score [UEMS], and quantitative grasping [QtG] measured at baseline.

The algorithm is based on two fundamental steps:

1. *Test of independence between endpoint and covariates, followed by variable selection*
   The statistical association of each possible covariate–endpoint pair is computed and a multiple-testing corrected p-value is assigned. If the global null hypothesis of total independence between covariates and outcome cannot be rejected, the algorithm stops without producing any split. Conversely, when at least one covariate is significantly associated with the clinical endpoint, then the algorithm selects the covariate with the strongest statistical association. The variable selected, as well as the p-value of its association with the outcome are reported in the inner nodes (circles in Figure 2).
2. *Splitting procedure*
   The algorithm evaluates all possible dichotomous splits on the covariate selected. The split is eventually set so as to maximize the discrepancy between the newly formed subgroups, making the subgroups as different as possible with regard to the endpoint selected. The best binary split is reported on the lines connecting nodes (see Figure 2).
3. *Recursive repetition*
   As long as the algorithm can identify covariates which are significantly associated to the endpoint, the previous steps of variable selection and splitting procedure are repeated in an iterative manner. When the algorithm stops, the outcome distribution in the terminal nodes is represented graphically (boxplots in Figure 2).

In contrast to other implementations of recursive partitioning, the separation of variable selection and splitting procedure prevents the systematic tendency towards covariate selection with many possible splits, independently of their association with the response (Hothorn *et al.*, 2006). In addition, it also permits the implementation of statistically motivated stopping criteria (Hothorn *et al.*, 2006). Recursive partitioning by conditional inference leans itself on a general recursive binary partitioning algorithm, but constructs independence tests under a permutation testing framework based on conditional distribution developed by Strasser and Weber (1999). Unbiased recursive partitioning by conditional inference is implemented by the function `ctree` of the add-on package **party** in the R system for statistical computing.

5.3 Advantages of the chosen approach

The application of unbiased recursive partitioning by conditional inference in the setting of spinal cord injury makes it possible to define relevant subgroups of patients at an early stage of a trial. This is easily achieved by training the regression model using covariates that correspond to clinical assessments performed shortly after injury. Consequently, as soon as baseline characteristics of patients are defined, data-driven, objective decision rules can be easily implemented to assess subgroup belonging and predicted recovery.

Subgroups defined by the algorithm can be applied both at the outset of a trial to define inclusion/exclusion criteria. In some instances, it may be advantageous to exclude a subgroup of subjects because their predicted spontaneous recovery is so minimal that even an effective treatment would not be detectable. On the other end of the spectrum, a subgroup of participants may be excluded because their predicted recovery is so extensive that ceiling effects would become a serious issue. The algorithm allows therefore a targeted enrolment of participants that not only are likely to respond to a treatment, but also where the treatment effect is likely to be detected.

For those participants who are included into the trial, the trial planner may resort to the partitioning model that already defined a stratification scheme. If the analyses are to be conducted strata-wise, our approach also offers clinicians with the option to define different, strata-specific endpoints before the study begins. Especially when considering activity-based endpoints, which are requested for confirmatory phase III trials, the definition of successful treatment effect may vary considerably among subgroups of participants with different baseline characteristics and different predicted recovery.

This new ability of being able to use broad inclusion criteria allows to efficiently allocate the few available participants (Wyndaele and Wyndaele, 2006), and at the same time define patient subpopulations which reduce the heterogeneity of spontaneous recovery patterns (Fawcett *et al.*, 2006). This peculiarity of the proposed approach is likely to prevent distortion of trial results due to a small number of participants exhibiting uncommon recovery patterns, as well as the detailed investigation of subtle treatment effects in more homogeneous subgroups. Undoubtedly, in addition to the exclusion of patient subgroups, a number of other criteria influences the final number of participants enrolled into a trial (the "funnel effect" in Jones *et al.* (2010)). Nonetheless, the potential inclusion of a large percentage of participants with different lesions combined with the a priori identification of more homogeneous subgroups represents a clear improvement in trial design.

Compared to clinical algorithms for the prediction of long-term outcomes and for patient stratification (Zörner *et al.*, 2010; van Middendorp *et al.*, 2011; Wilson *et al.*, 2012) our approach further provides a number of more technical advantages. Contrary to linear and logistic regression, unbiased recursive partitioning does not assume linear effects of the covariate (on

the link scale) and considers automatically interactions among covariates, providing a more flexible regression framework than usually applied in clinical research. In addition, the central step of variable selection is carried out on the p-value scale. The p-values for the test statistics can be directly compared, even if the covariates themselves are measured on different scales, which is a very likely scenario for all clinical settings. While linear or logistic regressions usually enforce a complete-case analysis with sometimes drastic consequences for the sample size, unbiased recursive partitioning prevents such large loss of data by momentarily excluding missing values in specific covariates during variable selection, but allowing measurements of a subject on all other covariates for the same algorithm step. In addition, our approach delivers a visualisation of the endpoint distribution of each identified subgroup, allowing a more direct examination and clinical validation of the obtained partitioning, as well as a visual aid to the formulation of possible subgroup-specific endpoints. More importantly, though, only unbiased recursive partitioning provides data-driven criteria for the identification of more homogeneous subgroups. Even the best-performing linear and logistic regression model leaves ultimately to the clinician the decision on where to draw the line for patient stratification, contributing therefore little to the main problem introduced by the necessity to design more inclusive trials.

6 Analysing a study: models for multiple-item rating scales

Endpoints generated by multiple-item rating scales are ubiquitous in neurological disciplines, and often represent the only data format which clinical assessments provide (Hobart, 2003). This is alike in the field of spinal cord injury, where most of motor, sensory, and functional outcomes are of this type (Catz *et al.*, 2007; Kirshblum *et al.*, 2011; Kalsi-Ryan *et al.*, 2012). The advent of Patient Reported Outcomes will further confront trial scientists with multiple-item rating scale outcomes. Nonetheless, despite the ubiquity of multiple-item rating scales used to measure health outcomes in patients with neurological disorders, current analysis approaches represent a major weakness of the translational process (Hobart *et al.*, 2007). In fact, ordinal scales are characterized by arbitrary numerical scores establishing a ranking of observations, where the difference between two following ranks is by no means bound to be equivalent across the entire item range. In addition, many neurological rating scales comprise items adressing and measuring several health domains. Therefore, the calculation of total summed scores and the use of statistical methods developed for continuous endpoints are questionable.

6.1 Current approaches

A historical revision of key studies in spinal cord injury (Lammertse, 2012) reveals that multiple-item rating scales are consistently analysed as a single overall summed score (or difference of total summed scores between baseline and follow-up) (Bracken *et al.*, 1984, 1990; Geisler *et al.*, 1991; Bracken *et al.*, 1997; Cardenas *et al.*, 2007). Nevertheless, those approaches have been shown to be inappropriate with regard to several aspects (Agresti, 2010), and consequences such as biased parameter estimates, misleading associations and loss of power are some of the known consequences of assuming metric properties for ordinal endpoints (Winship and Mare, 1984; Hastie *et al.*, 1989; Scott *et al.*, 1997).

6.2 Proportional odds model

To overcome previous limitations in the analysis of multiple-item rating scales such as the Upper Extremity Motor Scores (UEMS) in a classical two-armed randomized clinical trial, we proposed an autoregressive transitional ordinal model. The model is essentially a proportional odds regression model that includes baseline motor scores, a factor denoting the lesion level and the distance from it, an autoregressive term and a treatment indicator as explanatory variables.

The transitional ordinal model derives its name from the fact that this model analyses outcomes while controlling for the score obtained at the baseline assessment. The baseline motor scores are often included in trial analyses for baseline adjustment (Bracken *et al.*, 1997; Cardenas *et al.*, 2007), but their ordinality is discarded. The chosen modelling strategy is particularly useful in our setting: even in the case when the outcome distributions in the two trial arms were identical, the difference between baseline and end-of-trial assessment (the *transition*) may very well differ (Agresti, 2010).

The second term of the model codes for motor lesion level and the distance from it. The motor lesion level is introduced to ensure that only key muscles that have been affected by the injury, and on which an improvement of motor function can be actually recorded, are included into the analysis. Basically, the motor lesion level is defined as the most caudal key muscle with normal motor function. By definition, key muscles above the motor level have a maximal score on the rating scale. Since the vast majority of patients who suffered a spinal cord injury show a certain degree of recovery, the analysis of key muscles above the lesions is not sensible as they already have the maximal score. A further improvement, either due to spontaneous recovery or a treatment effect, cannot be recorded due to ceiling effects. Therefore, we analysed only key muscles below the motor level. Those key muscles have been affected by the injury, and any improvement of motor function can be readily recorded. In addition, we also coded for the distance of the key muscle being analysed to the motor level. This introduces a relabelling of the key muscles depending on their relative distance from the injury rather that their absolute position along the spinal cord, putting therefore all the affected key muscles on a common footing prior to analysis.

The autoregressive term of the model describes the anatomical structure of the spinal cord, and reproduces the decreasing pattern of motor scores with increasing distance from the motor level. Specifically, this term postulates that the motor score of a given key muscle depends explicitly on the motor score of the key muscle just rostral to it.

This formulation of the proposed transitional ordinal model includes both the most relevant prognostic factors such as baseline motor scores and motor lesion level, but also introduces a key muscle relabelling and interdependence derived from anatomical considerations.

Model fitting is implemented by the function `polr` of the add-on package **MASS** in the R system for statistical computing.

6.3 Latent variable model

Latent variable methods are successfully applied in situations where the variable of major interest cannot be measured directly, and must be inferred from a set of observed variables called indicators (Bartholomew *et al.*, 2008). A common illustrative example regards to the study of human intelligence: as this concept cannot be measured directly, psychologists collect information in the form of examination questions. The examination questions not only can be easily measured, but are also related to, or are an indicator of, the underlying concept of

interest (Everitt and Hothorn, 2011). The same logic applies to most primary endpoints in spinal cord injury: the key variable, which cannot be measured directly, is the neurological status of a patient. Instead, other measurable variables such as the Upper Extremity Motor Scores should be thought as a collection of indicators used to make inference about the latent variable of interest. This concept is rendered graphically as a path diagram in Figure 3.

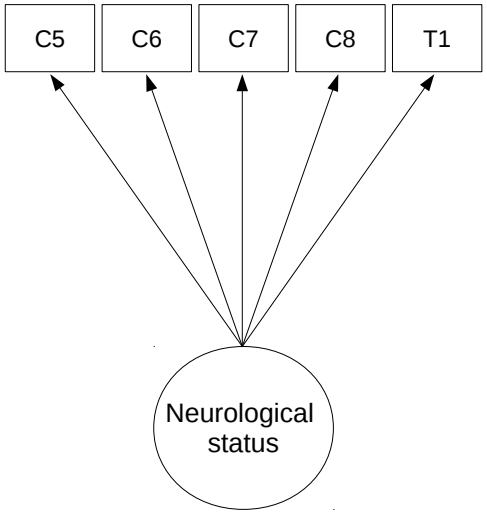

Figure 3.: Simplified path diagram representing the dependencies among key muscles as indicators (boxes), and the latent variable neurological status (circle).

In a setting where all indicators are measured on rating scales, and the latent variable inferred is metrical, the techniques are collectively referred to as Latent Trait Analysis (LTA). LTA relies on the assumption that, given the latent variabels, the indicators are independent from each other, or vice versa the correlations observed among the indicators are due to their relationship with a common latent trait. Under this postulate, Maximum Likelihood methods exist that choose the parameter values wich make the outcome frequency distribution predicted by the model as close as possible to the one observed in the data.

LTA models put the analysis of multiple-item rating scale on a much more rigorous and sophisticated inferential framework. It allows in fact to formally test the two major implicit assumptions of currently used analysis approaches, namely unit change and unidimensionality of the multiple-item rating scale. Firstly, a nominal LTA can shed light onto the validity of summing all motor scores in an unweighted total score. If all standardized loadings were of similar magnitude, this would imply a similar discriminating power of all items. As a similar weight would be applied to each response, the individual scores on the latent dimension would give a similar ranking as the total summed scores (Bartholomew *et al.*, 2008). Secondly, an ordinal LTA performed at single time points assesses whether all items or key muscles are indicators of a single, unidimensional latent health domain. This feature is tacitly assumed in current analysis approaches, but, to our knowledge, has never been tested formally. The dimensionality on a measuring instrument is assessed by fitting LTA models with increasing latent dimensions, which are then compared in terms of model fit to determine how many latent variables are necessary to best reproduce the observed data.

In addition, the longitudinal evolution of UEMS can also be modelled within the framework of latent variables. This essentially is achieved by adding an additional layer on top of the

LTA models. The first layer models the key muscles at each time point as indicators of the latent neurological status. The repeated measurement of the latent neurological status are thus themselves taken as indicators for so-called latent growth factors describing the intercept and slope components of the trajectory modelled (see Figure 4).

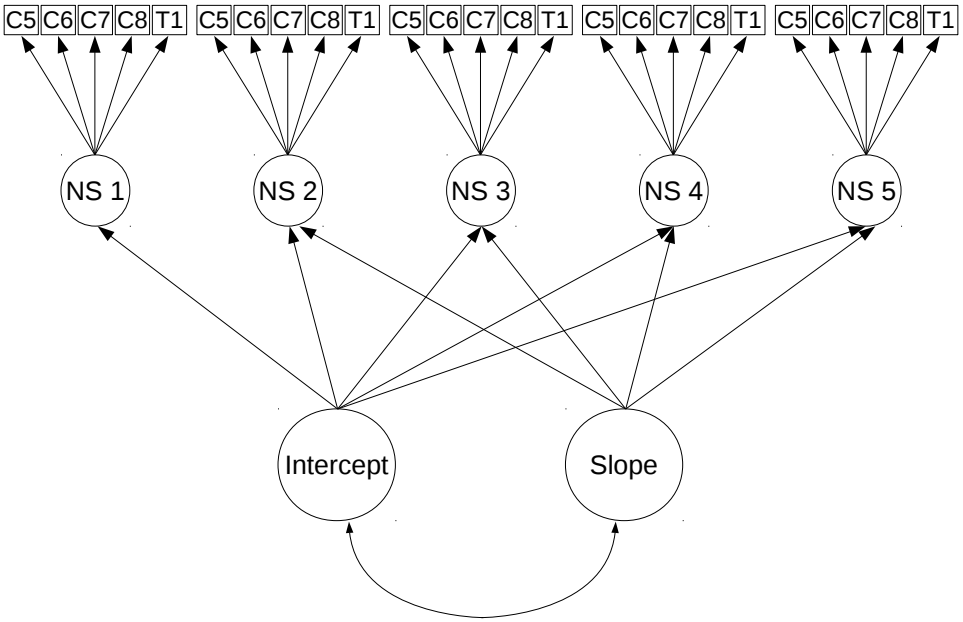

Figure 4.: Simplified path diagram representing the dependencies among the key muscles as indicators (boxes at the top), the latent variable neurological status at each time point (circles in the middle), which are themselves used as indicators of further latent variabels describing the intercept and slope component of the longitudinal growth modelling (circles at the bottom).

In a classical regression fashion, baseline covariates were allowed to directly influence both the intercept and slope components. In future studies, an indicator for treatment arm can be added in a straightforward manner, providing a test for treatment effect in the longitudinal analysis of neurological endpoints.

In the field of spinal cord injury, latent variable models of this type have been successfully used to investigate psychological traits (McColl *et al.*, 2003; Cruz-Almeida *et al.*, 2005; Martz *et al.*, 2005; Krause *et al.*, 2009; deRoon Cassini *et al.*, 2010; Catalano *et al.*, 2011; Bonanno *et al.*, 2012). Despite the similarity in scientific terms, to our knowledge, we are the first to apply the same methodology to a neurological endpoint, highlighting the shortcomings of currently adopted approaches.

Latent variable models are implemented in the statistical modeling program M*plus*.

7 Advantages of the chosen approach

For the first time in spinal cord injury, the analysis of the multiple-item rating scale of the Upper Extremity Motor Scores is carried out within an ordinal regression approach, putting its analysis on a much more appropriate and sophisticated framework. In fact, our approach prevents the known consequences of assuming continuous properties for summed multiple-item rating scores (Winship and Mare, 1984; Hastie *et al.*, 1989; Scott *et al.*, 1997; Agresti, 2010),

which is the current approach in the field.

The proposed transitional ordinal model analyses a sensible subset of key muscles (below the injury level), focusing therefore on those key muscles that were actually impacted by the lesion and can achieve a measurable recovery during rehabilitation. By including all the most relevant prognostic factors and a direct dependency on scores of adjacent key muscles, reproducing therefore the observed pattern of lower motor scores with increasing distance from the lesion, additional care is taken to make sure that the model is reflective of clinical and biological aspects of spinal cord anatomy and lesion severity.

The parameter that quantifies the treatment effect can be interpreted as a conditional odds ratio. In contrast to a clinically intangible improvement on mean motor scores provided by current approaches, odds ratios are a common and accepted way of quantifying treatment effect in the clinical setting and can be readily visualised in easily interpretable bar plots. Even in the case when the proportional odds assumption is not fully met, it still provides a meaningful summary of global treatment efficacy (Scott *et al.*, 1997).

Expanding on the important improvement provided by the proportional odds model, we proposed a strategy for the modelling of the longitudinal evolution of Upper Extremity Motor Scores. In essence, the latent variable model is a generalisation to more than two response categories of the standard logistic model for binary data, and is closely related to the proportional odds model. This approach is based on a statistical framework designed to model rating data with the specific structure of the longitudinal neurological score represented by UEMS. Not only is the ordinality of the data maintained, the approach also addresses in a formal way the assumptions of metric properties and dimensionality of current analysis approaches for spinal cord-specific multiple-item rating scales with the potential to invalidate analyses. Its flexibility further allows to incorporate an indicator for trial arm, providing a test for treatment effect in the longitudinal analysis of neurological endpoints.

Although the mathematical foundation of latent variable models is quite challenging (Hobart *et al.*, 2007), they have the potential to change the face of endpoint analysis in all health domains confronted with multiple-item rating scales (Hobart, 2003) by placing them within a much more rigorous and sophisticated inferential framework.

8 Outlook

Despite a much improved understanding of the mechanisms of neuronal plasticity and regeneration, clinical trials in neurological diseases are among the most challenging to conduct. In fact, a historical review of conducted spinal cord injury trials reveals that the majority of them did not succeed. As a consequence, there is still not an approved treatment for improving neurological function nor a consensus standard of care specific to people who suffered a spinal cord injury. The promises of preclinical discoveries have yet to be translated into treatments safely applicable to humans.

While the failure of any particular clinical trial can be imputed to a number of concomitant reasons, the poor choice of primary endpoints and their analysis as well as the large heterogeneity of few potential patients represent the hallmark of neurological trials. This thesis tackles both issues of patient inclusion and endpoint analysis, and proposes an innovative approach for the reliable stratification of homogeneous participant subgroups within a heterogeneous neurological disorder, as well as develops and tailors statistical methods used to analyse complex ordinal endpoints generated by multiple-item rating scales.

These new approaches and principles for decision-making in the clinical setting are an essential, yet previously missing component and will extend the current guidance concerning the conception of clinical trials. The latter had been condensed in four publications by the International Campaign for Cures of Spinal Cord Injury Paralysis, which touched on many crucial aspects such as trial design, inclusion and exclusion criteria, and clinical trial endpoints, but nonetheless in a rather descriptive nature. Our results provide indeed specific solutions for the planning of future clinical trials by providing data-driven decision rules for patient stratification and inclusion/exclusion as well as for the analysis of complex ordinal endpoints collected in spinal cord injury. At the same time, all the approaches presented are very flexible: the methodology can be applied to any particular set of clinical variables and patient population, and are therefore intended to act as a template across medical disciplines.

We are confident that the proposed approaches are going to be more and more applied in spinal cord injury and related disciplines, where the awareness of current limitations in the planning and analysis of trials has been rising. In this context, I think that future work and research will move on two parallel tracks.

On the one hand, clinicians and trial scientists will be confronted with the process of understanding, getting comfortable with, and finally adopting the proposed or similar approaches. Admittedly, the mathematical foundations and terminology of those methods can be quite challenging for researchers and clinicians with limited exposure to statistics. Nonetheless, those methods provide a huge potential to change the face of health outcomes by placing the analysis of complex ordinal endpoints within a much more rigorous and sophisticated inferential framework. I personally think that clinicians should strive to develop an understanding for those statistical methods, without delving into their mathematical foundations. Being able to use the methods to best serve their patients and research interests should be the goal to aim to, trusting the more technical issues to the method developers. The same approach is employed in the clinical setting for a number of other techniques routinely employed in hospitals such as rehabilitation robots and magnetic resonance imaging. While being very complex techniques, the accessibility (and not a detailed understanding) to the medical staff was one key element for their wide dissemination. In this sense, the diffusion

of such approaches could probably be greatly enhanced by developing and distributing some kind of down-loadable applications for smart phone and similar, as it has happened for many health and lifestyle monitoring devices. Being able to quickly and easily visualize several stratification scenarios with a few screen touches will greatly increase the attractiveness of the proposed approaches. To my knowledge, some effort in this direction is already underway.

On the other hand, statisticians will probably be looking at the combination of latent variables modeling strategies with recursive partitioning by conditional inference. In fact, our results clearly suggest that unbiased recursive partitioning is a very promising way to facilitate the design of more inclusive clinical trials. At the same time though, we reported clear evidence that the commonly adopted Motor Scores endpoint should not be analyses as a total sum of scores, but rather analysed within the framework of latent variable models, where each key muscle represent an indicator for a common underlaying health domain. I would therefore anticipate that more advanced approaches will first convert the Motor Scores into the corresponding position of a given patient on the latent variable, and subsequently this information will be fed into the recursive partitioning framework to obtain a partition of participants based on their latent neurological status. This approach could take advantage of both themes covered in this thesis and achieve a better patient stratification scheme preventing any inferential issues based on the inherent limitations of rating scales.

Future work and research on these two parallel tracks will be very important for the promotion of advanced and especially more appropriate statistics in the clinical field. A continuous exchange between statisticians and clinical scientists is going to be the key for a fruitful and successful interdisciplinary research stream delivering statistically improved approaches representing tailored clinical solutions. Ultimately, the ambition of all people involved and the results published within the framework of this thesis are intended to fast-forward the clinical process leading to specific therapeutic treatments for people living with spinal cord injury. We are glad that our proposed approaches are already being considered and employed in the planning and analysis of clinical trials!

Thesis Summary

This thesis consists of four papers. In pairs, the four projects address two major peculiarities of spinal cord injury trials, namely the high variability of the small patient population, and the analysis of neurological endpoints delivered by multiple-item rating scales. Both aspects are leading causes of the lagging translational process and require the application of more appropriate and innovative statistical methods.

Paper I and Paper II focus on the reliable stratification of homogeneous subgroups and the prediction of future clinical outcomes within a heterogeneous neurological disorder such as spinal cord injury. Paper I employed for the first time in the field the unbiased recursive partitioning technique called conditional inference tree, which is a tree-structured regression model based on sequential tests of independence specifically designed to identify more homogeneous subgroups within an initial heterogeneous patient population. Paper II provides a further application of unbiased recursive partitioning to a commonly targeted and clinically relevant population of participants to facilitate the design of more inclusive clinical trials. The comprehensive analyses provide a flexible template to help design future clinical studies.

Paper III and Paper IV address current limitations in the statistical analysis of neurological endpoints in the form of complex ordinal outcomes. In Paper III, we propose a transitional ordinal model with an autoregressive component in a common spinal cord trial setting. Paper IV discusses limitations of neurological rating scales and their longitudinal analysis within the unifying framework of latent variable modelling.

The content and contributions of each paper are briefly summarized below.

Paper I

Identifying Homogeneous Subgroups in Neurological Disorders: Unbiased Recursive Partitioning in Cervical Complete Spinal Cord Injury by Lorenzo G. Tanadini, John D. Steeves, Torsten Hothorn, Rainer Abel, Doris Maier, Martin Schubert, Norbert Weidner, Rüdiger Rupp, Armin Curt

The necessity to develop a strategy permitting the enrolment of participants with varying degrees of injuries to compensate for the limited pool of available patients (Sekhon and Fehlings, 2001), but also to maximise the scientific gain of knowledge for each completed trial can only be sensibly achieved with the capability to implement a stratification procedure to limit subject heterogeneity within study cohorts (Marino *et al.*, 1999; Tuszynski *et al.*, 2006).

This paper reports the first application of recursive partitioning by conditional inference (Hothorn *et al.*, 2006), a tree-based regression models specifically designed for screening heterogeneous populations and identifying more homogeneous subgroups. The paper introduces the model, and compares it to commonly adopted linear and logistic regression techniques (Zörner *et al.*, 2010; van Middendorp *et al.*, 2011; Wilson *et al.*, 2012). While confirming similar prediction accuracy for all approaches, a retrospective analysis of prospectively collected neurological data revealed the advantages of data-driven and easily implementable rational for early patient stratification provided by recursive partitioning.

The clinical setting and the necessity of reliable stratification of homogeneous cohorts were outlined by Prof. Curt and Prof. Steeves. The specific statistical approach of unbiased recursive partitioning by conditional inference was suggested by Prof. Hothorn. Data preparation, initial inference, as well as final analyses and the writing of the manuscript were done by

myself. Prof. Hothorn, Prof. Steeves, and Prof. Curt reviewed the manuscript at several stages and contributed to improve it. The remaining co-authors are members of the European Multicenter Study about Spinal Cord Injury, which collected and granted access to the data. They read and approved the final version of the paper. The period of my thesis in which this paper was produced was financially supported by the the International Foundation for Research in Paraplegia.

The main contribution of this paper is the context-specific presentation of a flexible approach for early patients stratification easily adaptable to future clinical studies confronted with small and/or heterogeneous patient populations.

Paper II

Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders: Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury Lorenzo G. Tanadini, Torsten Hothorn, Linda A. T. Jones, Daniel P. Lammertse, Rainer Abel, Doris Maier, Rüdiger Rupp, Norbert Weidner, Armin Curt, John D. Steeves

Building upon the results achieved in Paper I, this paper discusses the application of conditional inference trees on the clinically relevant group of patients with incomplete spinal cord injury. The inclusion of participants with incomplete lesions is justified, as many of the experimental approaches under scrutiny were developed using animal models with incomplete lesions. Given their spared function, it is also expected that patients with incomplete spinal cord injury are more likely to benefit from therapeutic interventions (Tuszynski *et al.*, 2006). However, the incomplete injury population is highly heterogeneous in terms of level and severity of injury, which gives rise to a diversity of recovery patterns (Marino *et al.*, 1999; Geisler *et al.*, 2001a; Fawcett *et al.*, 2006).

The paper applies conditional inference trees (Hothorn *et al.*, 2006) to a real-world trial setting based on retrospectively analysed, but prospectively collected neurological data, generating decision rules for the appropriate inclusion of participants. The prediction-based stratification of subjects with an incomplete lesion paves the way to a broad, but controlled inclusion and stratification of participants from a heterogeneous patient population. The algorithm also supports the selection of specific endpoints for each stratified cohort. The conditional inference tree produced was validated both internally as well as externally, providing stable and generalisable results.

The clinical setting and the research question were proposed by Prof. Curt and Prof. Steeves. The statistical approach of unbiased recursive partitioning was suggested by Prof. Hothorn. Data preparation, initial inference, as well as final analyses and the writing of the manuscript were done by myself. Prof. Hothorn, Prof. Steeves, Prof. Curt reviewed the manuscript at several stages and contributed to improve it. The remaining co-authors are members of the European Multicenter study about Spinal Cord Injury, which collected and granted access to the data. They read and approved the final version of the paper. The period of my thesis in which this paper was produced was financially supported by the the International Foundation for Research in Paraplegia.

The main contribution of this paper is the comprehensive analysis and validation of the clinically highly relevant cohort of incomplete spinal cord injury patients providing a template for a broad, but controlled inclusion and stratification of participants in future clinical trials.

Paper III

Autoregressive transitional ordinal model to test for treatment effect in neurological trials with complex endpoints by Lorenzo G. Tanadini, John D. Steeves, Armin Curt, Torsten Hothorn

Current approaches to the analysis of complex ordinal endpoints generated by multiple-item rating scales represent a major weakness of the translational process (Hobart, 2003; Hobart *et al.*, 2007). Although this type of endpoint is ubiquitous in spinal cord injury (Kirshblum *et al.*, 2011), routinely employed analysis approaches introduce several limitations (see Agresti (2010) for a general overview) by considering a single overall summed score and assuming it being a continuous measure (Bracken *et al.*, 1984, 1990; Geisler *et al.*, 1991; Bracken *et al.*, 1997; Cardenas *et al.*, 2007).

This paper proposes a transitional ordinal model with an autoregressive component to overcome widespread limitations in the analysis of endpoints in spinal cord injury (Winship and Mare, 1984; Hastie *et al.*, 1989; Scott *et al.*, 1997; Agresti, 2010; Hobart *et al.*, 2007). The model was fitted retrospectively to prospectively collected neurological data regarding a relevant primary endpoint in the field. A large simulation study of two-arm randomised clinical trial revealed an important gain in statistical power in virtually all simulation settings. A revisitation of a key historical trial (Geisler *et al.*, 1991) provides a head-to-head comparison between the proposed and currently employed analysis approaches, highlighting drawbacks of the latter.

The idea to resort to statistical approaches explicitly developed for the analysis of ordinal endpoints was suggested by Prof. Hothorn. In discussion with Prof. Hothorn, I developed the final model for the segment-wise analysis of the endpoint, which relies on the analysis of items (e.g. key muscles) depending on their distance along the spine from the site of injury. In addition, we included an autoregressive term to account for the anatomical structure of the spinal cord and the observed pattern of decreasing scores for more caudal segments. Data preparation, models fitting, implementation of the simulation, analysis of the historical trial, as well as the writing of the manuscript were done by myself. Prof. Hothorn reviewed the manuscript at several stages and provided support for the simulation implementation, Prof. Curt and Prof. Steeves provided inputs to an advanced version of the manuscript. This paper was partly produced with the financial support of the the International Foundation for Research in Paraplegia.

The main contribution of this paper is the implementation of an ordinal regression model which represents a specific solution to the analysis of a complex ordinal endpoint in spinal cord injury.

Paper IV

Addressing limitations of rating scales and their analysis in spinal cord injury under the unifying framework of latent variable modelling by Lorenzo G. Tanadini, Armin Curt, Irini Moustaki

Expanding on the important improvement provided by the ordinal approach, and supported by the specific structure of the longitudinal neurological data being analysed, we took full advantage of recent statistical developments in social statistics (Liu and Hedeker, 2006; Cagnone *et al.*, 2009; Vasdekis *et al.*, 2012), a field where the analysis of repeated ordinal measurements is common. In fact, modelling multivariate longitudinal ordinal responses based on a latent

variable approach seemed both appealing and appropriate for applications in the spinal cord setting.

This paper addresses metric properties and dimensionality of a spinal cord-specific multiple-item rating scale, and proposes an approach for the longitudinal modelling of neurological recovery following injury. The retrospective analysis of prospectively collected neurological data revealed that unidimensionality holds only for a subset of initially included patients. Important simplification in further analysis steps could be substantiated. The longitudinal modelling reported a strong negative correlation between intercept and slope terms.

I proposed the idea to apply statistical methods developed in the social sciences. In discussion with Prof. Moustaki, I finalised and implemented the assessment of metric properties and longitudinal modelling of the neurological endpoint chosen under the unifying framework of latent variable modelling. Data preparation, exploratory analyses, as well as final analyses and the writing of the manuscript were done by myself. Prof. Moustaki provided support during the implementation phase and reviewed the manuscript at several stages. Prof. Curt provided inputs to an advanced version of the manuscript. This paper was produced at the Department of Statistics of the London School of Economics, London UK with the financial support of a Swiss National Science Foundation doctoral mobility fellowship (Project P1ZHP3_158783) and the Janggen-Pöhn Foundation, St. Gallen.

The main contribution of this paper is the comprehensive analysis of complex ordinal endpoints entirely based on inferential framework of latent variable modelling.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*, Wiley Series in Probability and Statistics, second edn, Hoboken, New Jersey.

Andlin-Sobocki, P., Jönsson, B., Wittchen, H.-U. and Jes Olesen (2005). Cost of Disorders of the Brain in Europe, *European Journal of Neurology* **12**(Suppl. 1): 1–27.

Antonic, A., Sena, E. S., Lees, J. S., Wills, T. E., Skeers, P., Batchelor, P. E., Macleod, M. R. and Howells, D. W. (2013). Stem Cell Transplantation in Traumatic Spinal Cord Injury: A Systematic Review and Meta-Analysis of Animal Studies, *PLoS Biology* **11**(12): 1–14.

Bartholomew, D. J., Steele, F., Moustaki, I. and Galbraith, J. I. (2008). *Analysis of multivariate social science data*, Statistics in the Social and Behavioral Sciences, second edn, Chapman & Hall/CRC, London.

Bath, P. M. W., Gray, L. J., Collier, T., Pocock, S. and Carpenter, j. (2007). Can We Improve the Statistical Analysis of Stroke Trials? Statistical Reanalysis of Functional Outcomes in Stroke Trials, *Stroke* **38**(6): 1911–1915.

Bonanno, G. A., Kennedy, P., Galatzer-Levy, I. R., Lude, P. and Elfström, M. L. (2012). Trajectories of resilience, depression, and anxiety following spinal cord injury, *Rehabilitation Psychology* **57**(3): 236–247.

Bracken, M. B., Collins, W. F., Freeman, D. F., Shepard, M. J., Wagner, F. W., Silten, R. M., Hellenbrand, K. G., Ransohoff, J., Eisenberg, H. M., Rifkinson, N., Goodman, J. H., Meagher, J. N., Fischer, B., Clifton, G. L., Flamm, E. S. and Rawe, S. E. (1984). Efficacy of Methylprednisolone in Acute Spinal Cord Injury, *The Journal of the Americal Medical Association* **251**(1): 45–52.

Bracken, M. B., Shepard, M. J., Collins, W. F., Holford, T. R., Young, W., Baskin, D. S., Eisenberg, H. M., Flamm, E., Leo-Summers, L., Maroon, J., Marshall, L. F., Perot, P. L., Piepmeier, J., Sonntag, V. K., Wagner, F. C., Wilberger, J. E. and Winn, H. R. (1990). A randomized, Controlled Trial of Methylprednisolone or Naloxone in the Treatment of Acute Spinal-Cord Injury - Results of the Second National Acute Spinal Cord Injury Study, *The New England Journal of Medicine* **322**(20): 1405–1411.

Bracken, M. B., Shepard, M. J., Holford, T. R., Leo-Summers, L., Aldrich, E. F., Fazl, M., Fehlings, M., Herr, D. L., Hitchon, P. W., Marshall, L. F. and others (1997). Administration of methylprednisolone for 24 or 48 hours or tirilazad mesylate for 48 hours in the treatment of acute spinal cord injury: results of the Third National Acute Spinal Cord Injury Randomized Controlled Trial, *The Journal of the Americal Medical Association* **277**(20): 1597–1604.

Cagnone, S., Moustaki, I. and Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal responses, *British Journal of Mathematical and Statistical Psychology* **62**(2): 401–415.

Cardenas, D. D., Ditunno, J., Graziani, V., Jackson, A. B., Lammertse, D., Potter, P., Sipski, M., Cohen, R. and Blight, A. R. (2007). Phase 2 trial of sustained-release fampridine in chronic spinal cord injury, *Spinal Cord* **45**(2): 158–168.

Casha, S., Zygun, D., McGowan, M. D., Bains, I., Yong, V. W. and John Hurlbert, R. (2012). Results of a phase II placebo-controlled randomized trial of minocycline in acute spinal cord injury, *Brain* **135**(4): 1224–1236.

Catalano, D., Chan, F., Wilson, L., Chiu, C.-Y. and Muller, V. R. (2011). The buffering effect of resilience on depression among individuals with spinal cord injury: A structural equation model, *Rehabilitation Psychology* **56**(3): 200–211.

Catz, A., Itzkovich, M., Tesio, L., Biering-Sorensen, F., Weeks, C., Laramee, M. T., Craven, B. C., Tonack, M., Hitzig, S. L., Glaser, E., Zeilig, G., Aito, S., Scivoletto, G., Mecci, M., Chadwick, R. J., El Masry, W. S., Osman, A., Glass, C. A., Silva, P., Soni, B. M., Gardner, B. P., Savic, G., Bergström, E. M., Bluvshtein, V. and Ronen, J. (2007). A multicenter international study on the Spinal Cord Independence Measure, version III: Rasch psychometric validation, *Spinal Cord* **45**(4): 275–291.

Chow, A., Mayer, E. K., Darzi, A. W. and Athanasiou, T. (2009). Patient-reported outcome measures: The importance of patient satisfaction in surgery, *Surgery* **146**(3): 435–443.

Conigliani, C., Manca, A. and Tancredi, A. (2014). Prediction of patient-reported outcome measures via multivariate ordered probit models, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**(3): 567–591.

Cruz-Almeida, Y., Martinez-Arizala, A. and Widerstrm-Noga, E. G. (2005). Chronicity of pain associated with spinal cord injury: A longitudinal analysis, *The Journal of Rehabilitation Research and Development* **42**(5): 585.

De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models*, Statistics for Social Science and Public Policy, Springer, New York.

deRoon Cassini, T. A., Mancini, A. D., Rusch, M. D. and Bonanno, G. A. (2010). Psychopathology and resilience following traumatic injury: A latent growth mixture model analysis., *Rehabilitation Psychology* **55**(1): 1–11.

Dobkin, B., Apple, D., Barbeau, H., Basso, M., Behrman, A., Deforge, D., Ditunno, J., Dudley, G., Elashoff, R., Fugate, L., Harkema, S., Saulino, M., Scott, M. and SCILT Group (2006). Weight-supported treadmill vs over-ground training for walking after acute incomplete SCI, *Neurology* **66**(4): 484–493.

Everitt, B. and Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*, Use R!, Springer, New York.

Fawcett, J. W., Curt, A., Steeves, J. D., Coleman, W. P., Tuszynski, M. H., Lammertse, D., Bartlett, P. F., Blight, A. R., Dietz, V., Ditunno, J., Dobkin, B. H., Havton, L. A., Ellaway, P. H., Fehlings, M. G., Privat, A., Grossman, R., Guest, J. D., Kleitman, N., Nakamura, M., Gaviria, M. and Short, D. (2006). Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials, *Spinal cord* **45**(3): 190–205.

Fink, P., Ewald, H., Jensen, J., Sorensen, L., Engberg, M., Holm, M. and Munk-Jorgensen, P. (1999). Screening for somatization and hypochondriasis in primary care and neurologicl in-patients: a seven-item scale for hypochodriasis and somatization, **46**(3): 261–273.

Forrest, M. and Andersen, B. (1986). Ordinal scale and statistics in medical research., *British medical journal* **292**(6519): 537.

Geisler, F. H., Coleman, W. P., Grieco, G., Poonian, D. and Sygen Study Group (2001a). Measurements and recovery patterns in a multicenter study of acute spinal cord injury, *Spine* **26**(24S): S68–S86.

Geisler, F. H., Coleman, W. P., Grieco, G., Poonian, D. and Sygen Study Group (2001b). The Sygen® multicenter acute spinal cord injury study, *Spine* **26**(24S): S87–S98.

Geisler, F. H., Dorsey, F. C. and Coleman, W. P. (1991). Recovery of motor function after spinal-cord injury - A randomized, placebo-controlled trial with GM-1 ganglioside, *The New England Journal of Medicine* **324**(26): 1829–1838.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., Gannon, B., Jones, D. H., Jennum, P., Jordanova, A., Jönsson, L., Karampampa, K., Knapp, M., Kobelt, G., Kurth, T., Lieb, R., Linde, M., Ljungcrantz, C., Maercker, A., Melin, B., Moscarelli, M., Musayev, A., Norwood, F., Preisig, M., Pugliatti, M., Rehm, J., Salvador-Carulla, L., Schlehofer, B., Simon, R., Steinhausen, H.-C., Stovner, L. J., Vallat, J.-M., den Bergh, P. V., van Os, J., Vos, P., Xu, W., Wittchen, H.-U., Jönsson, B. and Olesen, J. (2011). Cost of disorders of the brain in Europe 2010, *European Neuropsychopharmacology* **21**(10): 718–779.

Hastie, T. J., Botha, J. L. and Schnitzler, C. M. (1989). Regression with an ordered categorical response, *Statistics in Medicine* **8**(7): 785–794.

Hawryluk, G. W., Rowland, J., Kwon, B. K. and Fehlings, M. G. (2008). Protection and repair of the injured spinal cord: a review of completed, ongoing, and planned clinical trials for acute spinal cord injury: A review, *Neurosurgical focus* **25**(5): 1–16.

Hobart, J. (2003). Rating scales for neurologists, *Journal of Neurology, Neurosurgery & Psychiatry* **74**(suppl 4): iv22–iv26.

Hobart, J. C., Cano, S. J., Zajicek, J. P. and Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations, *The Lancet Neurology* **6**(12): 1094–1105.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics* **15**(3): 651–674.

Jones, L. A. T., Lammertse, D. P., Charlifue, S. B., Kirshblum, S. C., Apple, D. F., Ragnarsson, K. T., Poonian, D., Betz, R. R., Knoller, N. and Heary, R. F. (2010). A phase 2 autologous cellular therapy trial in patients with acute, complete spinal cord injury: pragmatics, recruitment, and demographics, *Spinal cord* **48**(11): 798–807.

Kalsi-Ryan, S., Beaton, D., Curt, A., Duff, S., Popovic, M. R., Rudhe, C., Fehlings, M. G. and Verrier, M. C. (2012). The Graded Redefined Assessment of Strength Sensibility and Prehension: Reliability and Validity, *Journal of Neurotrauma* **29**(5): 905–914.

Kirshblum, S. C., Waring, W., Biering-Sorensen, F., Burns, S. P., Johansen, M., Schmidt-Read, M., Donovan, W., Graves, D. E., Jha, A., Jones, L., Mulcahey, M. J. and Krassioukov, A. (2011). Reference for the 2011 revision of the international standards for neurological classification of spinal cord injury, *The Journal of Spinal Cord Medicine* **34**(6): 547–554.

Kleitman, N. (2004). Keeping promises: translating basic research into new spinal cord injury therapies, *The journal of spinal cord medicine* **27**: 311–318.

Krause, J. S., McArdle, J. J., Pickelsimer, E. and reed, K. S. (2009). A Latent Variable Structural Path Model of Health Behaviors After Spinal Cord Injury, *The Journal of Spinal Cord Medicine* **32**(2): 162–174.

Kwon, B. K., Oxland, T. R. and Tetzlaff, W. (2002). Animal models used in spinal cord regeneration research, *Spine* **27**(14): 1504–1510.

Laffont, C. M., Vandemeulebroecke, M. and Concordet, D. (2014). Multivariate Analysis of Longitudinal Ordinal Data With Mixed Effects Models, With Application to Clinical Outcomes in Osteoarthritis, *Journal of the American Statistical Association* **109**(507): 955–966.

Lammertse, D. P. (2012). Clinical trials in spinal cord injury: lessons learned on the path to translation. The 2011 International Spinal Cord Society Sir Ludwig Guttmann Lecture, *Spinal cord* **51**(1): 2–9.

Lammertse, D. P., Jones, L. A. T., Charlifue, S. B., Kirshblum, S. C., Apple, D. F., Ragnarsson, K. T., Falci, S. P., Heary, R. F., Choudhri, T. F., Jenkins, A. L., Betz, R. R., Poonian, D., Cuthbert, J. P., Jha, A., Snyder, D. A. and Knoller, N. (2012). Autologous incubated macrophage therapy in acute, complete spinal cord injury: results of the phase 2 randomized controlled multicenter trial, *Spinal Cord* **50**(9): 661–671.

Lammertse, D., Tuszynski, M. H., Steeves, J. D., Curt, A., Fawcett, J. W., Rask, C., Ditunno, J. F., Fehlings, M. G., Guest, J. D., Ellaway, P. H., Kleitman, N., Blight, A. R., Dobkin, B. H., Grossman, R., Katoh, H., Privat, A. and Kalichman, M. (2006). Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: clinical trial design, *Spinal Cord* **45**(3): 232–242.

Liu, K., Tedeschi, A., Park, K. K. and He, Z. (2011). Neuronal Intrinsic Mechanisms of Axon Regeneration, *Annual Review of Neuroscience* **34**(1): 131–152.

Liu, L. C. and Hedeker, D. (2006). A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data, *Biometrics* **62**(1): 261–268.

Luther, S. L., Kromrey, J., Powell-Cope, G., Rosenberg, D., Nelson, A., Ahmed, S. and Quigley, P. (2006). A Pilot Study to Modify the SF-36v Physical Functioning Scale for Use With Veterans With Spinal Cord Injury, *Archives of Physical Medicine and Rehabilitation* **87**(8): 1059–1066.

Maas, A. I., Murray, G. D., Roozenbeek, B., Lingsma, H. F., Butcher, I., McHugh, G. S., Weir, J., Lu, J., Steyerberg, E. W. and IMPACT Study Group (2013). Advancing care for traumatic brain injury: findings from the IMPACT studies and perspectives on future research, *The Lancet Neurology* **12**(12): 1200–1210.

Marino, R. J., Ditunno, J. F., Donovan, W. H. and Maynard, F. (1999). Neurologic Recovery After Traumatic Spinal Cord Injury: Data From the Model Spinal Cord Injury Systems, *Archives of Physical Medicine and Rehabilitation* (80): 1391–1396.

Martz, E., Livneh, H., Priebe, M., Wuermser, L. A. and Ottomanelli, L. (2005). Predictors of Psychosocial Adaptation Among People With Spinal Cord Injury or Disorder, *Archives of Physical Medicine and Rehabilitation* **86**(6): 1182–1192.

McColl, M. A., Arnold, R., Charlifue, S., Glass, C., Savic, G. and Frankel, H. (2003). Aging, spinal cord injury, and quality of life: structural relationships, *Archives of Physical Medicine and Rehabilitation* **84**(8): 1137–1144.

McCullagh, P. (1980). Regression models for Ordinal Data, *Journal of the Royal Statistical Society* **42**(2): 109–142.

McHorney, C. A., Haley, S. M. and Ware, J. E. J. (1997). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative Precision Using Likert and Rasch Scoring Methods, *Journal of Clinical Epidemiology* **50**(4): 451–461.

Ravaud, J.-F., Delcey, M. and Yelnik, A. (1999). Construct validity of the functional independence measure (FIM): Questioning the unidimensionality of the scale and the" value" of FIM scores, *Scandinavian journal of rehabilitation medicine* **31**(1): 31–42.

Rudhe, C. and van Hedel, H. J. A. (2009). Upper Extremity Function in Persons with Tetraplegia: Relationships Between Strength, Capacity, and the Spinal Cord Independence Measure, *Neurorehabilitation and Neural Repair* **23**(5): 413–421.

Schwab, M. E. and Buchli, A. D. (2012). Drug research: plug the real brain drain, *Nature* **483**(7389): 267–268.

Scott, S. C., Goldberg, M. S. and Mayo, N. E. (1997). Statistical assessment of ordinal outcomes in comparative studies, *Journal of clinical epidemiology* **50**(1): 45–55.

Sekhon, L. H. and Fehlings, M. G. (2001). Epidemiology, demographics, and pathophysiology of acute spinal cord injury, *Spine* **26**(24S): S2–S12.

Sorani, M. D., Beattie, M. S. and Bresnahan, J. C. (2012). A Quantitative Analysis of Clinical Trial Designs in Spinal Cord Injury Based on ICCP Guidelines, *Journal of Neurotrauma* **29**(9): 1736–1746.

Steeves, J. D., Lammertse, D., Curt, A., Fawcett, J. W., Tuszynski, M. H., Ditunno, J. F., Ellaway, P. H., Fehlings, M. G., Guest, J. D. and Kleitman, N. (2006). Guidelines for the conduct of clinical trials for spinal cord injury (SCI) as developed by the ICCP panel: clinical trial outcome measures, *Spinal Cord* **45**(3): 206–221.

Strasser, H. and Weber, C. (1999). On the asymptotic theory of permutation statistics, *Mathematical Methods of Statistics* **8**: 220–250.

Tanadini, L. G., Hothorn, T., Jones, L. A., Lammertse, D. P., Abel, R., Maier, D., Rupp, R., Weidner, N., Curt, A. and Steeves, J. D. (2015). Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury, *Neurorehabilitation and Neural Repair* **29**(9): 867–877.

Tator, C. H. (2006). Review of treatment trials in human spinal cord injury: issues, difficulties, and recommendations, *Neurosurgery* **59**(5): 957–987.

Thuret, S., Moon, L. D. F. and Gage, F. H. (2006). Therapeutic interventions after spinal cord injury, *Nature Reviews Neuroscience* **7**(8): 628–643.

Tuszynski, M. H., Steeves, J. D., Fawcett, J. W., Lammertse, D., Kalichman, M., Rask, C., Curt, A., Ditunno, J. F., Fehlings, M. G. and Guest, J. D. (2006). Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP Panel: clinical trial inclusion/exclusion criteria and ethics, *Spinal Cord* **45**(3): 222–231.

Tutz, G. (2012). *Regression for categorical data*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, New York.

van Middendorp, J. J., Hosman, A. J., Donders, A. R. T., Pouw, M. H., Ditunno, J. F., Curt, A., Geurts, A. C., Van de Meent, H. and EMSCI Study Group (2011). A clinical prediction rule for ambulation outcomes after traumatic spinal cord injury: a longitudinal cohort study, *The Lancet* **377**(9770): 1004–1010.

Vasdekis, V. G., Cagnone, S. and Moustaki, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal responses, *Psychometrika* **77**(3): 425–441.

Velstra, I.-M., Bolliger, M., Tanadini, L. G., Baumberger, M., Abel, R., Rietman, J. S. and Curt, A. (2014). Prediction and Stratification of Upper Limb Function and Self-Care in Acute Cervical Spinal Cord Injury With the Graded Redefined Assessment of Strength, Sensibility, and Prehension (GRASSP), *Neurorehabilitation and Neural Repair* **28**(7): 632–642.

von Davier, M. and Carstensen, C. H. (2007). *Multivariate and Mixture Distribution Rasch Models*, Statistics for Social Science and Public Policy, Springer, New York.

Wahl, A. S., Omlor, W., Rubio, J. C., Chen, J. L., Zheng, H., Schroter, A., Gullo, M., Weinmann, O., Kobayashi, K., Helmchen, F., Ommer, B. and Schwab, M. E. (2014). Asynchronous therapy restores motor control by rewiring of the rat corticospinal tract after stroke, *Science* **344**(6189): 1250–1255.

Wilson, J. R., Grossman, R. G., Frankowski, R. F., Kiss, A., Davis, A. M., Kulkarni, A. V., Harrop, J. S., Aarabi, B., Vaccaro, A., Tator, C. H., Dvorak, M., Shaffrey, C. I., Harkema, S., Guest, J. D. and Fehlings, M. G. (2012). A Clinical Prediction Model for Long-Term Functional Outcome after Traumatic Spinal Cord Injury Based on Acute Clinical and Imaging Factors, *Journal of Neurotrauma* **29**(13): 2263–2271.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables, *American Sociological Review* **49**(4): 512–525.

Wyndaele, M. and Wyndaele, J.-J. (2006). Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey?, *Spinal cord* **44**(9): 523–529.

Zörner, B. and Schwab, M. E. (2010). Anti-Nogo on the go: from animal models to a clinical trial: Zörner & Schwab, *Annals of the New York Academy of Sciences* **1198**(Suppl.1): 22–34.

Zörner, B., Blanckenhorn, W. U., Dietz, V. and Curt, A. (2010). Clinical algorithm for improved prediction of ambulation and patient stratification after incomplete spinal cord injury, *Journal of neurotrauma* **27**(1): 241–252.

PAPER I

Identifying Homogeneous Subgroups in Neurological Disorders: Unbiased Recursive Partitioning in Cervical Complete Spinal Cord Injury

Lorenzo G. Tanadini, John D. Steeves, Torsten Hothorn, Rainer Abel, Doris Maier, Martin Schubert, Norbert Weidner, Rüdiger Rupp, Armin Curt

Paper published in *Neurorehabilitation and Neural Repair*, 2014, **Vol. 28** (6), 507-515.

Clinical Research Article

Identifying Homogeneous Subgroups in Neurological Disorders: Unbiased Recursive Partitioning in Cervical Complete Spinal Cord Injury

Neurorehabilitation and Neural Repair
2014, Vol. 28(6) 507–515

DOI: 10.1177/1545968313520413
nnr.sagepub.com
SAGE

Lorenzo G. Tanadini, MSc[1,2], John D. Steeves, PhD[3], Torsten Hothorn, PhD[2], Rainer Abel, MD[4,7], Doris Maier, MD[5,7], Martin Schubert, MD[1,7], Norbert Weidner, MD[6,7], Rüdiger Rupp, PhD[6,7], and Armin Curt, MD[1,7]

Abstract

Background. The reliable stratification of homogeneous subgroups and the prediction of future clinical outcomes within heterogeneous neurological disorders is a particularly challenging task. Nonetheless, it is essential for the implementation of targeted care and effective therapeutic interventions. *Objective*. This study was designed to assess the value of a recently developed regression tool from the family of unbiased recursive partitioning methods in comparison to established statistical approaches (eg, linear and logistic regression) for predicting clinical endpoints and for prospective patients' stratification for clinical trials. *Methods*. A retrospective, longitudinal analysis of prospectively collected neurological data from the European Multicenter study about Spinal Cord Injury (EMSCI) network was undertaken on C4-C6 cervical sensorimotor complete subjects. Predictors were based on a broad set of early (<2 weeks) clinical assessments. Endpoints were based on later clinical examinations of upper extremity motor scores and recovery of motor levels, at 6 and 12 months, respectively. Prediction accuracy for each statistical analysis was quantified by resampling techniques. *Results*. For all settings, overlapping confidence intervals indicated similar prediction accuracy of unbiased recursive partitioning to established statistical approaches. In addition, unbiased recursive partitioning provided a direct way of identification of more homogeneous subgroups. The partitioning is carried out in a data-driven manner, independently from a priori decisions or predefined thresholds. *Conclusion*. Unbiased recursive partitioning techniques may improve prediction of future clinical endpoints and the planning of future SCI clinical trials by providing easily implementable, data-driven rationales for early patient stratification based on simple decision rules and clinical read-outs.

Keywords

outcome prediction, clinical trial, cervical, sensorimotor complete, upper extremity motor score, motor level

Introduction

Traumatic spinal cord injury (SCI) is a heterogeneous disorder in terms of pathology, neurological deficits, and subsequent spontaneous recovery.[1,2] Furthermore, seemingly comparable cord injuries (as classified by the American Spinal Injury Association [ASIA] Impairment Scale [AIS A-E]), can achieve a diverse range of neurological and functional recovery, especially after incomplete SCI.[3] This is similar to other central nervous system disorders and makes reliable prediction of future outcomes challenging.

Despite this, the reliable prediction of future clinical endpoints is important to the implementation of targeted care and effective treatment options. In addition, reliably defining relatively homogeneous subgroups for clinical trials is

[1]Spinal Cord Injury Centre, Balgrist University Hospital, Zurich, Switzerland
[2]Division of Biostatistics, Institute for Social and Preventive Medicine, University of Zurich, Zurich, Switzerland
[3]ICORD, University of British Columbia and Vancouver Coastal Health, Vancouver, British Columbia, Canada
[4]Trauma Center Bayreuth, Bayreuth, Germany
[5]Trauma Center Murnau, Murnau, Germany
[6]Spinal Cord Unjury Center, Heidelberg University Hospital, Heidelberg, Germany
[7]EMSCI Study Group

Corresponding Author:
Lorenzo Tanadini, Spinal Cord Injury Center, Balgrist University Hospital, Forchstrasse 340, 8008 Zurich, Switzerland.
Email: ltanadini@paralab.balgrist.ch

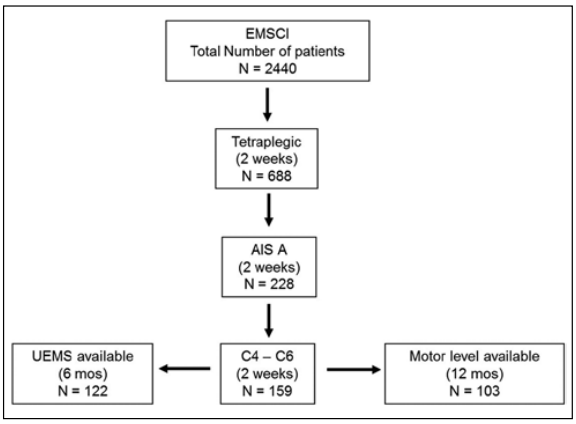

Figure 1. Subject numbers and selection criteria as extracted from the EMSCI database.
Abbreviations: EMSCI, European Multicenter study about Spinal Cord Injury; UEMS, Upper Extremity Motor Score; AIS, American Spinal Injury Association (ASIA) Impairment Scale.

important to accurately determining whether a therapeutic intervention provides a distinct benefit.[4-6] In many situations, such as a study of the biological or functional activity of an experimental therapeutic in early phase clinical trials, it is desirable to only enroll subjects who are relatively homogeneous in terms of both their early neurological status, as well as their prognosis for achieving a defined clinical endpoint (eg, future outcome). If trial participants have heterogeneous neurological and functional characteristics when assigned to a study arm, the contribution of a small number of participants may distort the overall results, the outcome interpretation, and disregard subtle treatment effects, thereby wasting subject and study resources.

Recently, in SCI, attempts have been made to create clinical algorithms for the prediction of long-term endpoints and for patient stratification.[7-9] These algorithms relied on the statistical techniques of multiple linear and logistic regressions. In this study, we compare these established statistical regression approaches[7-9] with a recently developed unbiased recursive partitioning regression tool called Conditional Inference Tree (URP-CTREE),[10] which directly identifies more homogeneous subgroups from an initial heterogeneous population. The aim was to compare the predictive accuracy of URP-CTREE against established regression models to predict future clinical endpoints from early neurological assessments, and to investigate the contribution of these methods to the stratification of cervical sensorimotor complete (AIS A) subjects into homogeneous subgroups.

Methods

Data Source

Data were obtained from the European Multicenter study on Spinal Cord Injury (EMSCI; http://www.emsci.org); an ongoing European network of SCI centers prospectively gathering data from subjects over the first year after traumatic SCI. The standardized assessment protocol tracks the neurological and functional status of patients during recovery from SCI. The EMSCI database was established in 2001 and has collected data from more than 2500 subjects during the past 12 years from 21 centers in 7 European countries.

Inclusion Criteria

The target population for this study included those EMSCI subjects who had cervical sensorimotor complete (AIS A) SCI with a Motor Level (from the right body side) at either C4, C5, or C6 as determined by a baseline assessment using the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI)[11] within the first 2 weeks after SCI (Figure 1). Only subjects with a documented assessment and determination of the selected clinical outcome of interest at that endpoint were included.

Two separate analyses are reported. The first analysis evaluated the total bilateral Upper Extremity Motor Score (UEMS) at 6 months after cervical complete SCI and is referred to hereafter as Total-UEMS. ISNCSCI motor score is determined by assigning to one muscle group, innervated and primarily identified with a specific spinal level, an integer between 0 (no detectable contraction) and 5 (active movement and a full range of movement against maximum resistance). Between C5 and T1 there are 5 representative "key" arm and hand muscles tested on each side of the body for a total upper extremity motor score of 25 + 25 = 50. The second analysis examined whether the subject achieved a 2-motor level improvement within the cervical cord (on either the left or right side) by 12 months after cervical complete SCI and is referred to as 2-motor level change. Subjects with a cervical complete SCI between C1 and C3 or C7 and T1 were excluded from analyses, as it is challenging to track the recovery of upper extremity motor scores or there were an insufficient number of subjects for statistical analysis.

Predictors and Clinical Endpoints

Potential clinical predictors (early ISNCSCI scores) and clinical endpoints (Total-UEMS and 2-motor level change at 6 and 12 months after SCI) were selected based on published literature[12-13] and clinical research experience of the authors. The set of predictors characterize the neurological status of the subjects according to the criteria of the ISNCSCI examination,[11] which is routinely assessed at all EMSCI specialized SCI care facilities within the first 2 weeks after injury (mean ± SD = 8.1 ± 4.7 days after injury). All predictors were collected according to EMSCI and ISNCSCI guidelines. Included predictors were age, the motor level (right body side), the bilateral sensory scores (light touch, pin prick), and motor scores (upper and lower extremity motor score), as well as information on the left

and right side for motor and sensory zone of partial preservation (ZPP) below the respective motor or sensory level. The zone of partial preservation refers to those segments caudal to the motor or sensory levels where there is some preservation of impaired motor or sensory function.

Here, we present 2 analyses based on complementary clinical endpoints. These endpoints have been related to determining changes in both neurological impairment and/or functional recovery (eg, independence in activities of daily living), as well as being suggested as possible clinical outcome measures for acute and/or subacute clinical studies involving cervical sensorimotor complete (AIS A) subjects.[14-16] Ancillary analysis for the bilateral total-UEMS at 12 months and the 2-motor level improvement within the cervical cord (on either the left or right side) by 6 months after cervical complete SCI were also performed.

Unbiased Recursive Partitioning: Conditional Inference Trees

The unbiased recursive partitioning technique called conditional inference tree (URP-CTREE) is a tree-structured regression model based on sequential tests of independence between predictors (eg, early clinical characteristics) and a specified clinical endpoint (ie, future outcome).[10] URP-CTREE divides an initial heterogeneous population into successively disjoint and more homogeneous pairs of subgroups with regard to the clinical endpoint of interest, and thus creates an algorithm for predicting future outcomes within more homogeneous subgroups.

URP-CTREE is based on 2 fundamental steps, which are repeated iteratively for each successive split of the initial heterogeneous population:

Step 1: Association of early predictors (subject's characteristics) with the clinical endpoint (outcome). The algorithm assesses whether any early predictor is statistically associated with the selected clinical endpoint. This is performed by individually calculating the statistical association of each possible predictor–endpoint pair (no data are presumed to be normally distributed). To each association, a multiple-testing corrected *P* value is assigned (ie, Bonferroni correction). If the initial null hypothesis of total independence between predictors and outcome cannot be rejected (no statistically significant association between any early predictor and the endpoint), the algorithm stops without producing any split of the initial population. On the contrary, if the null hypothesis of independence can be rejected, meaning that at least one early predictor is significantly associated with the subsequent clinical endpoint, then the algorithm selects the predictor with the strongest statistical association (smallest *P* value) and passes it to step 2.

Step 2: Splitting procedure for defining more homogeneous pairs of subgroups. Once the most significant predictor has been selected (as expressed in step 1), the algorithm evaluates all possible dichotomous splits on this variable, each one inevitably producing 2 subgroups. The goodness of each split is evaluated by a two-sample linear statistic (eg, χ^2 statistic for a binary outcome), to maximize the discrepancy between the newly formed subgroups. This partitions the initial population into 2 subgroups that are as distinct as possible.

Iterative steps: Recursively proceed to identify any additional early characteristics (predictors) that significantly predict the selected clinical endpoint. The recursive part of the algorithm starts over and the 2 fundamental steps (steps 1 and 2 listed above) are repeated separately for 2 newly formed subgroups. The URP-CTREE calculations proceed until no more statistically significant predictors are associated with the selected endpoint (null hypothesis cannot be rejected).

Once the clinical endpoint and predictors are selected, the algorithm will determine any significant associations without allowing any further input or bias by the investigator. Conditional inference trees can be applied to all types of regression problems, and has already been successfully used in other clinical settings with heterogeneous patient populations, like genetic marker–tumor association studies.[17,18]

Comparison of Statistical Methods

The recently developed URP-CTREE method is considered to directly identify more homogeneous study subgroups; however, its predictive accuracy needs to be compared against established statistical methods. Given the more continuous nature of the Total-UEMS endpoint, we compared multiple linear regressions, Least Absolute Shrinkage and Selection Operator (LASSO[19]) with URP-CTREE. Given the binary nature of the endpoint for a 2-motor level change, we compared multiple logistic regressions and LASSO with URP-CTREE. Linear and logistic regressions are well-known statistical techniques that have been previously employed in SCI research.[7-9] For the purpose of this article, LASSO can be interpreted as a multiple regression model with built-in variable selection.[19]

For evaluating the accuracy of Total-UEMS prediction, the models were compared by computing root mean square error (RMSE). RMSE is a frequently used measure of difference between observed values and values predicted by a model. It is defined as the root of the squared sum of differences between observed and predicted value divided by the total sample size.[20] The URP-CTREE–based prediction for continuous

Table 1. Root-mean-squared error as a measure of prediction accuracy for Total-UEMS at 6 months after cervical sensorimotor complete (AIS A) SCI. No statistically significant difference in accuracy between the three methods was observed.

	Multiple Regression	LASSO	Recursive Partitioning
95% CI lower bound	7.75	8.25	8.04
Median	11.02	10.41	10.36
95% CI upper bound	17.98	12.63	12.79

Abbreviations: 95% CI, 95% confidence interval.

Table 2. Misclassification rate as a measure of prediction accuracy for 2-motor level change at 12 months after cervical sensorimotor complete (AIS A) SCI. No statistically significant difference in accuracy between the three methods was observed.

	Logistic Regression	LASSO	Recursive Partitioning
95% CI lower bound	11.3	14.5	12.5
Median	22.5	26.3	24.4
95% CI upper bound	36.6	38.2	37.7

Abbreviations: 95% CI, 95% confidence interval.

outcomes is computed as the final node–specific mean. For assessing the accuracy of 2-motor level change prediction, the models were compared by computing the misclassification rate. Misclassification rate for a binary outcome is defined as the percentage of incorrect future outcome prediction based on the model, compared with the actually observed values.[20] The URP-CTREE–based prediction for binary outcomes is based on the final node–specific most likely outcome.

Following standard benchmarking procedures,[21] both measures were based on 500 bootstrap iterations. All analyses were performed in the computing environment R,[22] version 2.14.0, and based on the package party: A Laboratory for Recursive Partitioning.[23]

Results

Comparison of Statistical Methods for Predicting Clinical Endpoints

The prediction accuracy (RMSE) of Total-UEMS at 6 months after SCI is based on 500 bootstrap iterations and shown in Table 1. All 3 statistical approaches provide similar median and overlapping 95% confidence intervals (CIs) for RMSE.

Likewise, the examined statistical methods for predicting a 2-motor level change on either side of the cervical cord within the first 12 months after cervical complete (AIS A) SCI also provide similar statistical accuracy (Table 2). The difference here is the selected clinical endpoint was a binary event. Based on 500 bootstrap iterations, Table 2 shows the 95% CI for misclassification rate. All 3 statistical approaches provide similar median and overlapping 95% CI.

Ancillary analyses for Total-UEMS at 12 months and 2-ML recovery at 6 months provided similar results in terms of median and CI across the different methods. Therefore, for reasons of clarity, this article refers to the primary analysis only.

Direct Identification of More Homogeneous Study Subgroups

Figure 2 shows the conditional inference tree (URP-CTREE) for Total-UEMS at 6 months after cervical sensorimotor complete (AIS A) SCI.

In the example shown in Figure 2, the iterative identification of more homogeneous subgroups is based on 2 significant early neurological predictors, namely UEMS and Motor ZPP, as measured within the first 2 weeks after cervical sensorimotor complete SCI. Successive UEMS or ZPP cutoff values are indicated at the "branch points" 1, 2, 4, and 7. At each branch point, a multiple testing–adjusted P value is given, which describes the strength of the statistical association between the early predictor (UEMS or ZPP) and the endpoint (Total-UEMS). The full distribution of the Total-UEMS is revealed in the box plots within the final nodes at the bottom. The visual representation of a conditional tree is directly interpretable and easily applied in a clinical setting to determine which subjects to include, exclude, or group together for a desired study. Conversely, conventional linear regression methods will deliver an equation of the type

$$\begin{aligned}\text{Outcome} = \text{Intercept} + \\ \text{Slope}_1 * \text{Predictor}_1 + \ldots + \\ \text{Slope}_n * \text{Predictor}_n\end{aligned}$$

which quantifies how predictors are associated with the chosen endpoint, but does not provide direct decision rules for stratification of subjects into more homogeneous study subgroups.

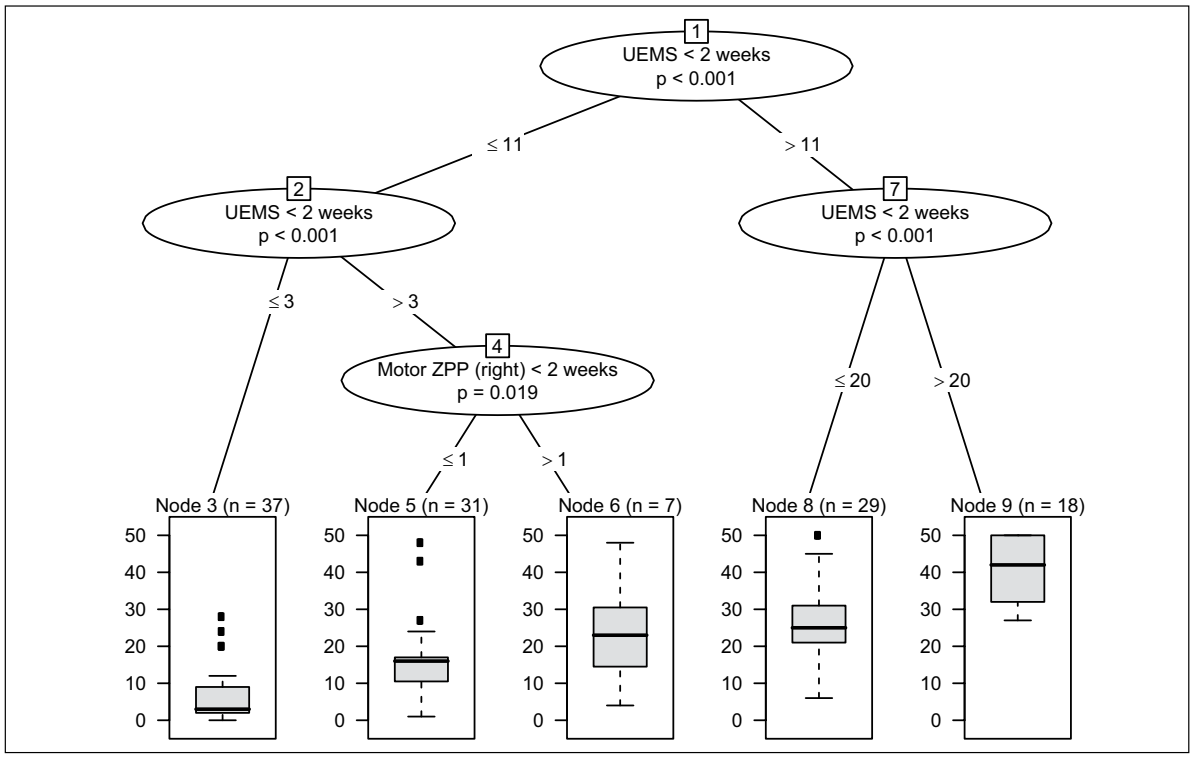

Figure 2. Conditional inference tree for the endpoint Total-UEMS at 6 months after cervical sensorimotor complete (AIS A) SCI (N=122), using a broad set of neurological and functional predictors assessed within the first two weeks after injury. The upper part represents the sequential splits based on early predictors (nodes 1,2,4,7); the lower part represents the achieved partition of the initial population into 5 more homogeneous subgroups, as represented by the final nodes (nodes 3, 5, 6, 8, 9). Boxplots at the bottom show sample size and distribution of the clinical endpoint within each subgroup (node).

Figure 3 shows the conditional inference tree (URP-CTREE) for the 2-motor level change within 12 months after cervical complete (AIS A) SCI. The URP-CTREE algorithm led to a partition of the initial EMSCI cervical AIS A population into 2 subgroups based on UEMS (as measured within 2 weeks after injury) and described in the final nodes (nodes 2 and 3). A multiple testing–adjusted *P* value is given, which describes the strength of the statistical association between the early predictor characteristic (UEMS) and the endpoint (2-motor level change). The full distribution of the clinical endpoint is revealed within nodes 2 and 3. Once again, the visual representation of a conditional tree is directly interpretable and can be implemented in a clinical setting. Conventional logistic regression methods deliver equation of the type

$$\text{Logarithm}(\text{Prob}_{\text{success}} / \text{Prob}_{\text{failure}}) = \text{Intercept} + \text{Slope}_1 * \text{Predictor}_1 + \ldots + \text{Slope}_n * \text{Predictor}_n$$

and do not provide direct decision rules for stratification.

Discussion

Clinical prediction models for the prognosis of potential future outcomes as well as for the identification of subgroups of SCI patients having predictable recovery patterns are essential.[1,4,24] Several attempts have been made to create clinical algorithms for the prediction of future clinical endpoints and for patient stratification in SCI.[7-9] These attempts rely on the statistical techniques of multiple linear regressions and logistic regression (for which the following considerations also apply). Despite achieving in some cases excellent discrimination in prediction of future clinical endpoints,[7,8] these approaches present shortcomings (see Table 3 for an overview) that may have hindered their wider application in SCI. Here, we present unbiased recursive partitioning's conditional inference trees (URP-CTREE) as a statistical method that overcomes some of these challenges.

Comparable Prediction Accuracy

In an attempt to overcome some of the drawbacks of established regression models, we applied the statistical method

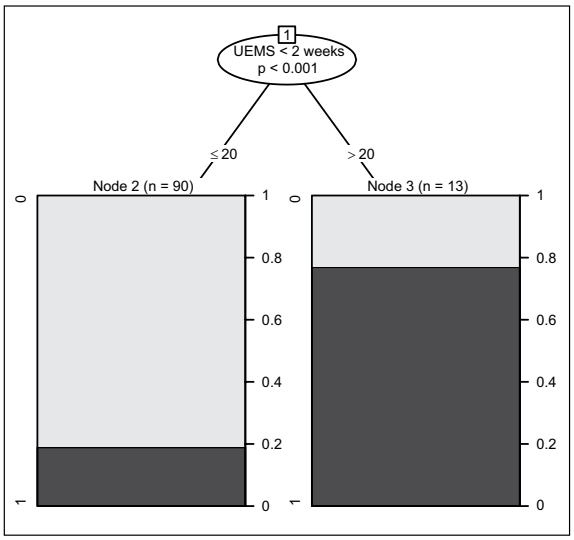

Figure 3. Conditional inference tree for the recovering of at least two motor levels (black shading; grey shading indicating failure of achieving this specified endpoint) at 12 months after cervical sensorimotor complete (AIS A) SCI (N=103), using a broad set of neurological and functional predictors assessed within the first two weeks after injury. The upper part represents the splits based on early predictors (here only node 1); the lower part represents the achieved partition of the initial population into 2 more homogeneous subgroups, as represented by the final nodes (nodes 2, 3). Plots at the bottom show sample size and distribution of the clinical endpoint for each subgroup.

of conditional inference trees from the family of unbiased recursive partitioning methods (URP-CTREE) for the first time in SCI. As a first prerogative for its consideration, we established that URP-CTREE provides equal statistical accuracy in predicting selected clinical endpoints (future outcomes) from a broad set of early clinical characteristics taken from neurological and functional assessments across a sample of cervical sensorimotor complete SCI subjects (Tables 1 and 2). In both analyses, median estimates of accuracy are similar for all 3 methods tested. Confidence intervals based on resampling techniques clearly indicate no statistical differences in accuracy across statistical methods. In general terms, there is no consensus on how to define standard reference values for accuracy (eg, correlation coefficient and the differentiation between weak, moderate, and strong); it has to be evaluated depending on the specific setting of application.

Drawbacks of Established Regression Methods

Even though linear models are powerful statistical tools they may lack specific information that may be essential for clinical applications. Multiple linear regression quantifies how a given set of early predictors associates with the mean of a future clinical endpoint and provides a numeric equation of these relationships. Linear predictor–outcome relationships are tacitly assumed and statistical interactions rarely modeled (Table 3). Especially in complex neurobiological settings, the assumption of a strictly linear relationship between predictors and clinical outcome, with no interactions between predictors do not seem sensible.[1,15,16,25] In addition, focusing on just parameter estimation (eg, the mean) prevents an understanding of the full endpoint distribution within a study population (for which the mean is only its central tendency).

Even more importantly in the context of clinical trials, neither linear nor logistic regression provide a direct and objective mean of partitioning an initial, heterogeneous population into more homogeneous subgroups, leaving the need for stratification unmet. For example, a fitted linear regression model provides as a mathematical equation (see Results section) quantifying the relationships between predictors and endpoints. This equation still leaves challenges on its implementation in clinical settings, where the handling of such equations cannot be easily implemented to inform about the prognosis for an individual subject (Table 3). In addition, multiple regression showed a higher upper bound for the confidence interval (Table 1). This is likely to be due to collinearity, a situation which arises when different predictors are highly correlated, causing difficulties in model fitting and interpretation. Collinearity is an issue in established regression techniques, but prevented in URP-CTREE (and LASSO).

Conditional Inference Trees

In contrast to established regression methods, URP-CTREE does not assume linear dependence between predictors and endpoint, and it specifically puts the modelling focus on interactions between predictors[10] (Table 3). In addition, URP-CTREE has the major advantage of defining more homogeneous subgroups in a direct, data-driven manner and to reveal the clinical endpoint distribution within each subgroup[10] (Table 3). These unique features of URP-CTREE could be of value for refining the stratification of patients for future clinical trials. URP-CTREE could also be used as an explorative tool for defining sensible primary and secondary outcomes for specific subgroups.

Total-UEMS and 2-Motor Level Analyses

The specific advantages of URP-CTREE outlined above are clearly visible in the results provided by the 2 analyses performed. Figure 2 represents the application of URP-CTREE to the clinical endpoint of Total-UEMS at 6 months after SCI. The occurrence of subsequent splits along the same UEMS scale within Figure 2 strongly suggests that it cannot

Table 3. Summary of key differences between unbiased recursive partitioning (URP-CTREE) and multivariate linear/logistic regression models.

	Linear/Logistic regression	Unbiased Recursive Partitioning
Clinical endpoint (outcome)	Choice of investigator	Choice of investigator
Early predictors (eg, neurological or functional)	Choice of investigator	Choice of investigator
Linear effects	Assumed	Not assumed
Interactions among predictors	Usually not considered	Considered
Missing values in predictors	Not allowed	Allowed
Scaling of predictors	Difficult to interpret	Accounted for
Endpoint considerations	Average (mean) only	Full distribution
Homogeneous cohort (inclusion/exclusion criteria)	Not directly	Directly

be assumed that the recovery between the baseline and 6 months occurs as a linear function. This is not a new finding[1,15,16,25] but stresses the importance of nonlinear effects and interactions, which are readily detected by URP-CTREE.[10] However, routine regression analyses tacitly assume linearity and usually do not account for interactions. Figure 2 also clearly suggests that even within the narrowly defined patient subgroups of cervical sensorimotor complete (AIS A) there will be variability in recovery, underscoring the limited value of an AIS grade both as a stratification tool as well as its change as a sensitive measure for any subtle or meaningful therapeutic effect.[1,16,25] In our sample of sensorimotor complete SCI subjects, URP-CTREE also provides a way of identifying subjects subgroups that will potentially show flooring (Figure 2, node 3) and ceiling (Figure 2, node 9) effects, which is of great relevance for the planning of clinical trials and the definition of primary endpoints.

Figure 3 represents the application of URP-CTREE to a binary clinical endpoint, 2-motor level change within 1 year after injury. The analysis suggests that cervical AIS A subjects with initial UEMS less or equal 20 (node 2, Figure 3) have a much lower probability of spontaneous recovery of at least 2 motor levels than subjects with a higher initial UEMS (node 3, Figure 3). URP-CTREE shows that selecting subjects only from node 2 of Figure 3 would provide even more homogeneous subgroups than the inclusion of all cervical sensorimotor complete subjects,[16] which would translate in a lower false positive rate for a possible clinical study based on such outcome. Nevertheless, the purpose of the 2-motor level change analysis was to demonstrate that URP-CTREE works for different types of clinical endpoints (continuous Total-UEMS and binary 2-motor level change) measured at 2 different time points (6 and 12 months after injury), and this partitioning within the population of cervical complete SCI may not always be necessary or preferred. In fact, the overall percentage of subjects recovering at least 2 motor levels within 1 year after injury (32/117; 27%) agrees with previous findings,[16] and has been suggested as a clinically meaningful primary outcome for cervical sensorimotor complete (AIS A) SCI population as a whole.

Homogeneous Subgroups

Our results indicates that URP-CTREE, while being specifically designed to identify more homogeneous subgroups within an initially heterogeneous population, does not compromise prediction accuracy when compared to established statistical regression approaches. Notably, the same conclusion is reached for 2 different clinical endpoints (continuous Total-UEMS change and binary 2-motor level change) measured at 2 different time points (Figures 2 and 3). Ancillary analysis based on Total-UEMS after 12 months and 2-motor level change after 6 months provided similar results (similar medians and overlapping confidence intervals) and further evidence for our conclusions.

We based our analyses on a sensorimotor-complete SCI population, which is clinically usually recognized as a rather homogeneous population in the context of SCI. Our analyses show that even within the narrowly defined patient subgroups of cervical sensorimotor complete (AIS A), there will be substantial variability in recovery (Figure 2), suggesting the need for a more differentiated approach. Nonetheless, we recognize that the full potential of URP-CTREE is expected to be realized in even more heterogeneous population, for example, individuals living with incomplete SCI.

Choosing Endpoints and Predictors

As with every statistical modeling approach, including URP-CTREE, the choice of the clinical endpoint is central because it directly influences (*a*) which predictors are significantly associated with it (step 1 of URP-CTREE), (*b*) where the dichotomous splits are set (step 2 of URP-CTREE), and (*c*) how resulting subgroups (nodes) are defined. In short, choose your clinical endpoint to assure it is appropriate to the "target" of your therapeutic intervention. The same cannot be said for early predictors; any number of reasonable data traits can be entered into URP-CTREE with the only consequence of making the correction for multiple testing (eg, Bonferroni) more stringent, but

URP-CTREE will still only identify those predictors that significantly associate with the chosen clinical endpoint. URP-CTREE can handle all types of predictors and several types of clinical endpoints.[10]

Limitations

Linear and logistic regression models are most useful when the relationship between early predictors and the clinical endpoint under investigation is truly linear, but this has not been demonstrated for SCI[1] or any other central nervous system disorder. In settings where this assumption holds, established regression methods are likely to outperform URP-CTREE in terms of prediction accuracy.

The present study was designed to introduce URP-CTREE and assess its value for predicting future clinical endpoints and stratifying heterogeneous populations. While the resampling technique confirms the validity of our conclusions, generalizability of URP-CTREE trees shown here should be evaluated in independent samples of subjects.

Many clinical assessments like UEMS are analyzed as sum scores of different items and often treated as continuous variables for further analysis, even though they are ordinal scales. We acknowledge that this could provide misleading results if sum scores do not represent a consistent scoring metric. Rasch analysis could provide insight into the measurement properties of commonly used clinical assessments[26] and produce a measurement scale that can be more confidently analyzed, but this is beyond the scope of the present study.

Conclusion

The results of our analysis show that URP-CTREE provides advantages over established multivariate linear and logistic regression techniques without compromising prediction accuracy. Above all, conditional inference trees are specifically designed to identify more homogeneous subgroups within an initial heterogeneous patient population. Data-driven, objective decision rules for more homogeneous subgroup identification can be created and easily implemented in clinical studies. URP-CTREE can be applied to all kind of regression problems, and could therefore be applied to a wide range of neurological disorders where the identification of more homogeneous subgroups is desired.

Acknowledgments

The authors acknowledge the support of the European Multicenter study about Spinal Cord Injury network (EMSCI), the International Foundation for Research in Paraplegia (IFP), the Spinal Cord Outcomes Partnership Endeavor (SCOPE), and the Clinical Research Priority Program in Neuro-rehabilitation of the University of Zurich. We appreciated the constructive comments of Linda Jones on an early draft of this article and the continuous assistance of René Koller with the EMSCI database.

Authors' Note

Rainer Abel, Doris Maier, Martin Schubert, Norbert Weidner, Rüdiger Rupp, and Armin Curt are members of the EMSCI (European Multicenter study about Spinal Cord Injury) Study Group.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Fawcett JW, Curt A, Steeves JD, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. *Spinal Cord*. 2006;45:190-205.
2. Bracken MB, Holford TR. Neurological and functional status 1 year after acute spinal cord injury: estimates of functional recovery in National Acute Spinal Cord Injury Study II from results modeled in National Acute Spinal Cord Injury Study III. *J Neurosurg*. 2002;96(3 suppl):259-266.
3. Marino RJ, Ditunno JF Jr, Donovan WH, Maynard F Jr. Neurologic recovery after traumatic spinal cord injury: data from the Model Spinal Cord Injury Systems. *Arch Phys Med Rehabil*. 1999;80:1391-1396.
4. Tuszynski MH, Steeves JD, Fawcett JW, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP Panel: clinical trial inclusion/exclusion criteria and ethics. *Spinal Cord*. 2006;45:222-231.
5. Casha S, Zygun D, McGowan MD, Bains I, Yong VW, Hurlbert RJ. Results of a phase II placebo-controlled randomized trial of minocycline in acute spinal cord injury. *Brain*. 2012;135(pt 4):1224–1236.
6. Lammertse DP, Jones LA, Charlifue SB, et al. Autologous incubated macrophage therapy in acute, complete spinal cord injury: results of the phase 2 randomized controlled multicenter trial. *Spinal Cord*. 2012;50:661-671.
7. Wilson JR, Grossman RG, Frankowski RF, et al. A clinical prediction model for long-term functional outcome after traumatic spinal cord injury based on acute clinical and imaging factors. *J Neurotrauma*. 2012;29:2263-2271.
8. van Middendorp JJ, Hosman AJ, Donders AR, et al; EM-SCI Study Group. A clinical prediction rule for ambulation outcomes after traumatic spinal cord injury: a longitudinal cohort study. *Lancet*. 2011;377:1004-1010.
9. Zörner B, Blanckenhorn WU, Dietz V, Curt A. Clinical algorithm for improved prediction of ambulation and patient stratification after incomplete spinal cord injury. *J Neurotrauma*. 2010;27:241-252.
10. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat*. 2006;15:651-674.

11. Kirshblum SC, Waring W, Biering-Sorensen F, et al. Reference for the 2011 revision of the International Standards for Neurological Classification of Spinal Cord Injury. *J Spinal Cord Med*. 2011;34:547-554.
12. Al-Habib AF, Attabib N, Ball J, Bajammal S, Casha S, Hurlbert RJ. Clinical predictors of recovery after blunt spinal cord trauma: systematic review. *J Neurotrauma*. 2011;28:1431-1443.
13. Wilson JR, Cadotte DW, Fehlings MG. Clinical predictors of neurological outcome, functional status, and survival after traumatic spinal cord injury: a systematic review. *J Neurosurg*. 2012;17:11-26.
14. Rudhe C, van Hedel HJ. Upper extremity function in persons with tetraplegia: relationships between strength, capacity, and the spinal cord independence measure. *Neurorehabil Neural Repair*. 2009;23:413-421.
15. Steeves JD, Lammertse DP, Kramer JL, et al. Outcome measures for acute/subacute cervical sensorimotor complete (AIS-A) spinal cord injury during a phase 2 clinical trial. *Top Spinal Cord Inj Rehabil*. 2012;18:1-14.
16. Kramer JL, Lammertse DP, Schubert M, Curt A, Steeves JD. Relationship between motor recovery and independence after sensorimotor-complete cervical spinal cord injury. *Neurorehabil Neural Repair*. 2012;26:1064-1071.
17. Owzar K. Alternate statistical tools and limitations in genetic marker association studies in single-arm drug cancer trials. *J Clin Oncol*. 2008;26:1400-1401.
18. De Roock W, Claes B, Bernasconi D, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol*. 2010;11:753-762.
19. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc*. 1996;58:268-288.
20. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY: Springer; 2013.
21. Hothorn T, Leisch F, Zeileis A, Hornik K. The design and analysis of benchmark experiments. *J Comput Graph Stat*. 2005;14:675-699.
22. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2011. http://www.R-project.org/.
23. Hothorn T, Hornik K, Strobl C, et al. party: a laboratory for recursive partitioning. R package. 2013. version 1.0-9. http://CRAN.R-project.org/package=party
24. Steeves JD, Lammertse D, Curt A, et al. Guidelines for the conduct of clinical trials for spinal cord injury (SCI) as developed by the ICCP panel: clinical trial outcome measures. *Spinal Cord*. 2006;45:206-221.
25. Steeves JD, Kramer JK, Fawcett JW, et al. Extent of spontaneous motor recovery after traumatic cervical sensorimotor complete spinal cord injury. *Spinal Cord*. 2010;49:257-265.
26. Catz A, Itzkovich M, Tesio L, et al. A multicenter international study on the Spinal Cord Independence Measure, version III: Rasch psychometric validation. *Spinal Cord*. 2006;45:275-291.

PAPER II

Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders: Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury

Lorenzo G. Tanadini, Torsten Hothorn, Linda A. T. Jones, Daniel P. Lammertse, Rainer Abel, Doris Maier, Rüdiger Rupp, Norbert Weidner, Armin Curt, John D. Steeves

Paper published in *Neurorehabilitation and Neural Repair*, 2015, **Vol. 29** (9), 867-877.

Clinical Research Article

Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders: Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury

Neurorehabilitation and Neural Repair
2015, Vol. 29(9) 867–877

DOI: 10.1177/1545968315570322
nnr.sagepub.com
SAGE

Lorenzo G. Tanadini, MSc[1,2], Torsten Hothorn, PhD[2], Linda A. T. Jones, MSc[3], Daniel P. Lammertse, MD[4,5], Rainer Abel, MD[6,7], Doris Maier, MD[7,8], Rüdiger Rupp, PhD[7,9], Norbert Weidner, MD[7,9], Armin Curt, MD[1,7], and John D. Steeves, PhD[10]

Abstract

Background. Several novel drug- and cell-based potential therapies for spinal cord injury (SCI) have either been applied or will be considered for future clinical trials. Limitations on the number of eligible patients require trials be undertaken in a highly efficient and effective manner. However, this is particularly challenging when people living with incomplete SCI (iSCI) represent a very heterogeneous population in terms of recovery patterns and can improve spontaneously over the first year after injury. *Objective*. The current study addresses 2 requirements for designing SCI trials: first, enrollment of as many eligible participants as possible; second, refined stratification of participants into homogeneous cohorts from a *heterogeneous* iSCI population. *Methods*. This is a retrospective, longitudinal analysis of prospectively collected SCI data from the European Multicenter study about Spinal Cord Injury (EMSCI). We applied conditional inference trees to provide a prediction-based stratification algorithm that could be used to generate decision rules for the appropriate inclusion of iSCI participants to a trial. *Results*. Based on baseline clinical assessments and a defined subsequent clinical endpoint, conditional inference trees partitioned iSCI participants into more homogeneous groups with regard to the illustrative endpoint, upper extremity motor score. Assuming a continuous endpoint, the conditional inference tree was validated both internally as well as externally, providing stable and generalizable results. *Conclusion*. The application of conditional inference trees is feasible for iSCI participants and provides easily implementable, prediction-based decision rules for inclusion and stratification. This algorithm could be utilized to model various trial endpoints and outcome thresholds.

Keywords

unbiased recursive partitioning, inclusion exclusion criteria, upper extremity motor score, clinical trial design

Introduction

In the field of spinal cord injury (SCI) research, many discoveries from basic research and clinical investigation are worthy of consideration as potential SCI therapeutics and have entered the translational process (for an overview of SCI clinical trials, see www.scope-sci.org). As SCI has a relatively low incidence,[1] there is a limited number of potentially appropriate acute and subacute trial participants. With the increasing number of new interventions, the recruitment of participants to acute and subacute SCI trials needs to be developed in a more inclusive and effective manner.

If SCI trials only proceed by sequentially enrolling participants with very narrow inclusion criteria,[2] so as to ensure subject homogeneity, enrollment progress will remain slow, affecting negatively the timely and financially sustainable accomplishment of the planned goals. In addition, with a

[1]Spinal Cord Injury Center, Balgrist University Hospital, Zurich, Switzerland
[2]Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[3]Craig H. Neilsen Foundation, Encino, CA, USA
[4]Craig Hospital, Englewood, CO, USA
[5]Department of Physical Medicine and Rehabilitation, University of Colorado Denver, Aurora, CO, USA
[6]Trauma Center Bayreuth, Bayreuth, Germany
[7]EMSCI Study Group
[8]Trauma Center Murnau, Murnau, Germany
[9]Spinal Cord Injury Center, Heidelberg University Hospital, Heidelberg, Germany
[10]ICORD, University of British Columbia and Vancouver Coastal Health, Vancouver, Canada

Corresponding Author:
Lorenzo Tanadini, Spinal Cord Injury Center, Balgrist University Hospital, Forchstrasse 340, Zurich 8008, Switzerland.
Email: ltanadini@paralab.balgrist.ch

narrow recruitment strategy, the scientific gain of knowledge brought about by the completion of a trial will be restricted to this narrow SCI subpopulation, requiring additional studies for SCI subjects with other characteristics and being therefore inefficient. Alternatively, once safety has been established, a more inclusive strategy would enroll participants with varying degrees of sensorimotor complete and incomplete SCI (iSCI). However, this can only be sensibly implemented with stratification algorithms to limit subject heterogeneity within a study cohort.[3,4]

Including iSCI participants is justified and required since many of the experimental treatments being translated for human SCI were developed using animal models with incomplete lesions. It is also expected that iSCI trial participants are more likely to benefit from interventions that target the spared sensorimotor function. An inclusive enrollment strategy will more efficiently utilize potential trial participants, as well as decrease the time and therefore the costs for completing a SCI trial program.

However, inclusive recruitment and enrollment strategies of iSCI participants poses specific challenges. Incomplete SCI comprises a highly heterogeneous population in terms of level and severity of injury, as well as the diversity of recovery patterns.[4] Since iSCI participants have some preserved sensory and/or motor function, any adverse events could also compromise spontaneous recovery.

Predictive algorithms are required that will quickly and accurately exclude those patients whose spontaneous recovery would be so extensive it would obscure any therapeutic benefit. In an acute study, the algorithm would have to accurately stratify participants on the basis of early clinical characteristics. Thus, the selection of sensitive and reliable endpoints is critical and such endpoints may be specific to a subset of stratified cohorts.

To capitalize on an inclusive recruitment approach, while carefully monitoring its limitations, requires the ability to recognize early clinically relevant subgroups of patients. The idea of identifying and stratifying similar study participants is demanding, but not new.[5] In this study, we propose an advanced prediction-based stratification approach based on a recently developed Unbiased Recursive Partitioning technique called Conditional inference TREEs[6] (URP-CTREE). Our aim is to assess the potential of URP-CTREE in defining a priori stratifications in a heterogeneous neurological disorder as represented by iSCI participants.

Methods

Data Source

The data utilized in this study was extracted from the European Multicenter study about Spinal Cord Injury (EMSCI, www.emsci.org, ClinicalTrials.gov Identifier: NCT01571531) according to the inclusion criteria reported below. EMSCI encompasses 21 centers from 7 European countries that have been tracking the functional and neurological status of patients during the first year of recovery from SCI in a rigorously standardized way since 2001.

Inclusion Criteria

From the 2597 patients present in the EMSCI database, we extracted all those de-identified individuals with an initial baseline assessment (<2 weeks) having the most caudal spinal segment with intact sensorimotor function (ie, neurological level of injury [NLI]) between C4 and C7 as defined by the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI)[7] and described by the American Spinal Injury Association Impairment Scale (AIS). SCI participants have typically been classified using the 5 grades (A-E) of the AIS. These SCI severity categories are broad and have limited value for research studies. Nevertheless, it is a commonly used classification system and a reasonable reference point for describing the participants in this study. We included AIS-B iSCI participants with sensory perception to S4 to S5, but no motor function preserved more than 3 segments below the cervical NLI (sensory incomplete SCI), as well as AIS-C participants (nonfunctional motor incomplete), where sensory and motor function are partially preserved below the NLI, but more than half the key muscles, below the NLI, have a muscle grade strength of <3/5. An additional minimum criterion was a neurological assessment at 6 months with entry of an upper extremity motor score (UEMS), since this was the primary endpoint of this illustrative analysis. Patients with suspected central cord syndrome[8] were excluded (n = 5), as routinely done in many clinical trials.

We also excluded from this analysis tetraplegic sensorimotor complete (AIS-A) patients, as they have been previously examined and reported.[9] Finally, we excluded cervical AIS-D participants. Spontaneous motor recovery after a cervical AIS-D injury is substantial with little room on the UEMS scale to measure a therapeutic benefit (ie, ceiling effect).

Predictors and Clinical Endpoint

Clinical neurological predictors included in this analysis, as well as the illustrative clinical endpoint, were selected based on the published literature[10,11] and clinical experience of the authors. The set of clinical characteristics (predictors) characterizing the neurological status of patients at baseline (assessed <2 weeks after iSCI) included the following: the NLI, the right and left motor levels, the right and left sensory levels, the bilateral sensory scores (light touch, pin prick), and motor scores (upper and lower extremity motor score), as well as epidemiological information, such as age

at injury and cause of injury. The set of predictors characterize the neurological status of the subjects according to the criteria of the ISNCSCI examination.[7]

The illustrative clinical endpoint chosen for this study was the bilateral UEMS at 6 months after injury, which consists of the sum of the manual muscle tested (MMT) strength for each of the 10 key muscles of the upper extremity (5 on each side of the body) and ranges between 0 and 50. This endpoint has been previously related to changes in upper extremity function,[12] as well as suggested[13] as potential endpoint in clinical trials. A number of trials also considered the change in UEMS as primary endpoint.[3,14] In this regard, we note that the stratification based on our illustrative endpoint does not preclude the use of change in motor score as trial endpoint. In fact, the 2 endpoints are strongly interconnected and their simultaneous considerations seems sensible. First, the direct application of URP-CTREE on the change in UEMS creates a situation for which both participants with severe lesions and those with high initial UEMS achieve similar, small change in UEMS (the first due to the severity of their lesion, the latter solely due to the ceiling effect on the scale used to measure motor function), making stratification for trials questionable. In addition, one has to consider UEMS so as to know whether the postulated treatment effect (especially if defined on the change in UEMS) can actually be measured on the UEMS scale. It should be also noted that, since the identification of more homogeneous groups with regard to UEMS at 6 months after injury is based on baseline UEMS (among others), information on changes in UEMS is actually contained in our conditional tree (see Figure 1). More generally, UEMS is an impairment-based clinical endpoint and may or may not be the most appropriate primary endpoint for a future trial. Phase III pivotal trials will likely require a function-based endpoint. Functional (activity-based) endpoints can be analyzed using the methods described here, but this is beyond the scope of this article.

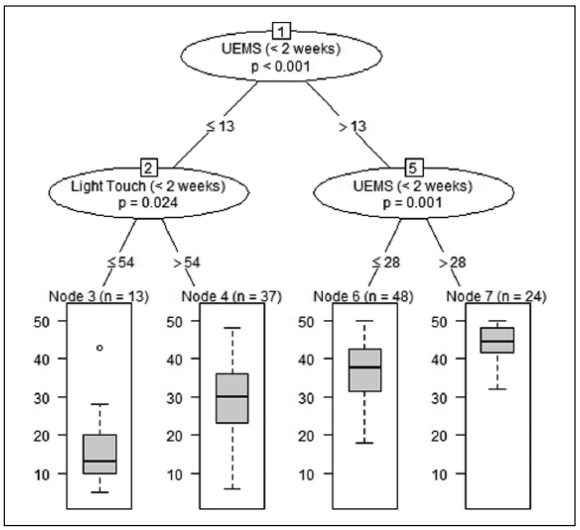

Figure 1. Conditional inference tree for the endpoint Upper Extremity Motor Score (UEMS) at 6 months after cervical motor-incomplete (American Spinal Injury Association Impairment Scale B and C) spinal cord injury (N = 122), using a broad set of neurological predictors assessed within the first 2 weeks after injury.

The algorithm led to a partition of the initial patient population into 4 terminal nodes. Each inner node (nodes 1, 2, 5) presents a cut-point into the initial, heterogeneous population and reports the variable selected with multiple testing-corrected *P* value (significance level is *P* = .05) and the split points on the "branches." The first split separates patients with an initial UEMS ≤13 or >13. Further partitioning is based on Light Touch (≤54 or >54) and UEMS (≤28 or >28). Terminal nodes (nodes 3, 4, 6, 7) display the identified subgroups. Boxplots at the bottom show sample size and distribution of the illustrative endpoint UEMS at 6 months after injury, defined as the sum of the manual muscle tested (MMT) strength for each of the 10 key muscles of the upper extremity (5 on each side of the body), and ranging 0 to 50. For more details on the interpretation of the conditional inference tree, please refer to the Methods section.

Conditional Inference Trees

In our analysis, we adopted the approach of recursive partitioning by conditional inference[6] (URP-CTREE). Tree-based regression models of this type are particularly useful for screening heterogeneous patient populations to identify more homogeneous patient subgroups. In fact, this type of approach has been successfully applied in disparate clinical fields, including cardiovascular disease,[15] genetic marker–tumor association studies,[16,17] and traumatic brain injury.[18]

The URP-CTREE algorithm is based on the iterative application of a variable selection and a population splitting procedure. In the first step, the algorithm assesses whether any early clinical predictor (baseline variable) is associated with the selected clinical endpoint. To each association, a multiple testing-corrected *P* value based on permutation tests is assigned. When no early clinical predictor conveys any information about the clinical endpoint (response variable), and therefore no association is statistically significant ($P \geq .05$), the algorithm stops. In contrast, when at least one early predictor is significantly associated with the selected clinical endpoint, a split in the patient population is performed, based on the predictor that showed the most significant association with the endpoint. The split is established in a way that maximizes the difference between the 2 newly formed subgroups with regard to the endpoint. Subgroups that will be further split in successive iterations are referred to as inner nodes and shown graphically as circles indicating the predictor on which a further split is performed (as well as the multiple-testing corrected *P* value for the predictor-endpoint association; see upper part of Figure 1). In that sense, inner nodes are just graphical placeholders for subgroups that

will be further split by the algorithm. The steps outlined above are reiterated until there are no statistically significant associations left between early clinical predictors and the selected clinical endpoint. When this is the case, the algorithm stops and returns a structure resembling an upside-down tree that divides an initially heterogeneous study population into more homogeneous subgroups with respect to the endpoint specified. Figure 1 shows an example where the initial, heterogeneous iSCI population "enters" the tree on the top inner node and is subsequently subdivided by the internal rules of the algorithm till no further partitioning is possible. At this stage, the endpoint distribution within each subgroup is displayed in the URP-CTREE's terminal nodes (see boxplots in Figure 1).

A detailed technical description with several applications of the statistical model is reported in Hothorn et al,[6] and a more pertinent SCI explanation, using cervical AIS-A SCI, has been recently completed.[19] In this previous study,[19] we investigated the value of URP-CTREE in comparison to established multivariate statistical approaches (eg, linear and logistic regression) for predicting clinical endpoints and for prospective stratification of sensorimotor-complete (AIS-A) SCI participants. Our analyses concluded that, while all methods employed had similar prediction accuracy, URP-CTREE provided additional advantages by directly identifying more homogeneous subgroups of patients, and describing the full endpoint distribution within each subgroup.

Fitting and Validation

Following model fitting (see Figure 1) with the EMSCI data, we inspected the resulting conditional inference tree with regard to variable selection and split point estimation. Internal validation of the EMSCI-based URP-CTREE was investigated by means of a resampling technique. From the original EMSCI sample, we drew 1000 random samples with replacement. For each random sample (of sample size n = 122), we fitted a new conditional inference tree allowing the same predictors. We then investigated the frequency of predictor selection and the distribution of split points.

External validity was assessed using a similar but independent sample of patient data from the Sygen[20] trial. We applied the decision rules provided by the inner nodes of the EMSCI-based tree (Figure 1) to Sygen participants with initial characteristics complying with the inclusion criteria outlined above. We then compared EMSCI-based terminal nodes to the results obtained for Sygen participants. Ninety-five percent confidence intervals for the median difference in all 4 pairs of terminal nodes were computed. All analysis were performed in the computing environment R, version 3.0.1,[21] and based on the statistical packages "party: A laboratory for Recursive Partitioning (Version 1.0-13)"[22] for conditional inference trees, and "coin: Conditional Inference Procedures in a Permutation Test Framework (Version 1.0-23)"[23] for confidence interval computation.

Table 1. Distribution of the Illustrative Endpoint UEMS at 6 Months After Injury for Each Terminal Node in the Conditional Inference Tree Based on EMSCI Participants[a].

UEMS	Node 3	Node 4	Node 6	Node 7
Minimum	5	6	18	32
25% Quantile	10	23	31.75	41.75
Median	13	30	37.5	44.5
75% Quantile	20	36	42.25	48
Maximum	43	48	50	50
AIS-B (total 53)	8 (15%)	18 (34%)	15 (28%)	12 (23%)
AIS-C (total 69)	5 (7%)	19 (27%)	33 (48%)	12 (17%)

Abbreviations: UEMS, upper extremity motor score; EMSCI, European Multicenter study about Spinal Cord Injury; AIS, American Spinal Injury Association Impairment Scale.
[a]Baseline ASIA Impairment Scale grades distribution are also reported.

Results

Fitting the Conditional Inference Tree

URP-CTREE for UEMS at 6 months based on a sample of 122 tetraplegic (ie, cervical) AIS-B (n = 53) and AIS-C (n = 69) participants extracted from EMSCI is shown in Figure 1. The iterative identification of more homogeneous subgroups is based on 2 baseline neurological assessments, namely, UEMS and light touch sensory perception. Splits are indicated at the inner nodes 1, 2, 5. URP-CTREE produced 4 terminal nodes (3, 4, 6, 7), for which the endpoint distribution is shown in the boxplots at the bottom. Table 1 reports a numeric summary of outcome distribution as well as the initial baseline AIS distribution for each terminal node (Figure 1).

Variable Selection

Variable selection is the fundamental step of URP-CTREE and relies on the computation of permutation tests between predictors and endpoint as well as the identification of the most significant associations. The *P*-value-based ranking of the best 4 predictors for each inner node is reported in Table 2. Baseline UEMS is the variable selected in inner node 1 (*P* value < .001) and inner node 5 (*P* value = .001), and baseline light touch sensation is selected in inner node 2 (*P* value = .024). The EMSCI-based conditional inference tree shows a stable variable selection, for which the selected variable is clearly the most significant from within all predictors included in the analysis. The conditional inference tree stopped growing (no further split was implemented), as no further predictor–endpoint association was significant at the *P* value = .05 level.

Table 2. Variable Selection for the Inner Nodes of the EMSCI-Based Conditional Inference Tree[a].

Node 1		Node 2		Node 5	
Predictors	*P* Value	Predictors	*P* Value	Predictors	*P* Value
UEMS	3.1×10^{-11}	Light touch	.024	UEMS	.001
ML (right)	.0002	UEMS	.437	ML (right)	.410
ML (left)	.001	ML (right)	.542	ML (left)	.476
NLI	.001	Pin prick	.868	Age	.724
Other predictors		Other predictors		Other predictors	

Abbreviations: EMSCI, European Multicenter study about Spinal Cord Injury; UEMS, upper extremity motor score; ML, motor level; NLI, neurological level of injury.
[a]The first 4 predictors are reported by increasing *P* value (decreasing importance). Nominal significance value is $P = .05$.

AIS Grades

As can be seen in Table 1, the initial classification of iSCI severity as AIS-B or AIS-C has little predictive value for the distribution of UEMS scores at 6 months. If it did, then most of the AIS-B participants would be stratified to cohort 3 where the lowest final UEMS scores were recorded and AIS-C participants would be more evident within cohort 7 where the highest UEMS scores were observed. Instead, a participant initially classified AIS-B was just as likely to show a substantial amount of motor function as an AIS-C participant (terminal node 7 has 12 AIS-B and 12 AIS-C). Thus, UEMS after 6 months could not be predicted by the initial baseline AIS grade. AIS-B and AIS-C participants at 6 months were distributed across all cohorts, often fairly equally.

Split Points

Following variable selection, a split is set so as to maximize discrepancy between the 2 newly formed subgroups. The split points estimated in each inner node are shown in Figure 2. The plots represent the standardized test statistics[6] for each possible split as a measure of subgroup discrepancy; the vertical dotted line is the split point used in the EMSCI-based URP-CTREE. All vertical dotted lines aligned with the value that maximizes discrepancy between the 2 newly formed subgroups. This substantiated the validity of all 3 split points for the EMSCI-based conditional inference model. Even for inner node 2, whose split statistics seem to have 2 local maxima (2 "bumps"), the split was produced at the highest value statistic available and was therefore correctly implemented.

Internal Validation

We examined the internal validity of the EMSCI-based tree using a standard resampling technique and investigated the frequency of predictor selection and the distribution of split points. Results of this resampling technique are shown in Figure 3. For inner node 1, baseline UEMS was selected in all resampling iterations (predictor 8, top panel left, Figure 3), and the split points resulted in an almost symmetric distribution around the first split point defining the maximal discrepancy for the 2 subgroups in EMSCI (vertical line, top panel right).

For inner node 2, baseline light touch was consistently selected, and no pattern of alternative variable selection emerged (predictor 6, middle panel left, Figure 3). For all iterations where light touch was the variable selected, the second split point distribution showed an approximately symmetric distribution centered at the split point implemented in the EMSCI-based conditional inference tree (vertical line, middle panel right).

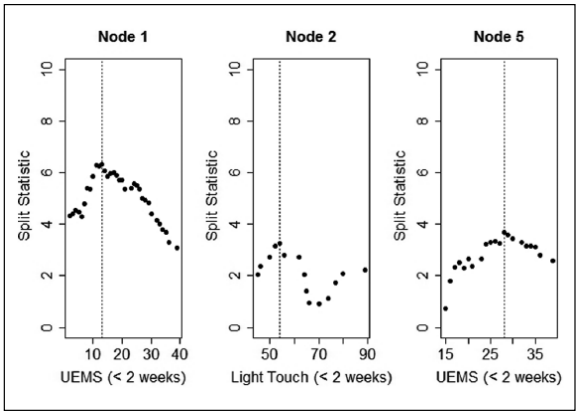

Figure 2. Split point estimation of inner nodes for the European Multicenter Study about Spinal Cord Injury-based conditional inference tree.
Each dot represents a possible partitioning of the patient population based on the most significant predictor (early participant characteristic). The split statistic is a measure of discrepancy for the 2 newly formed subgroups and is maximized by the algorithm such that the split partitions the patient population in subgroups that are as different as possible. The vertical dotted line represents the split point implemented. For more details on the interpretation of the conditional inference tree, please refer to the Methods section.
Abbreviation: UEMS, upper extremity motor score.

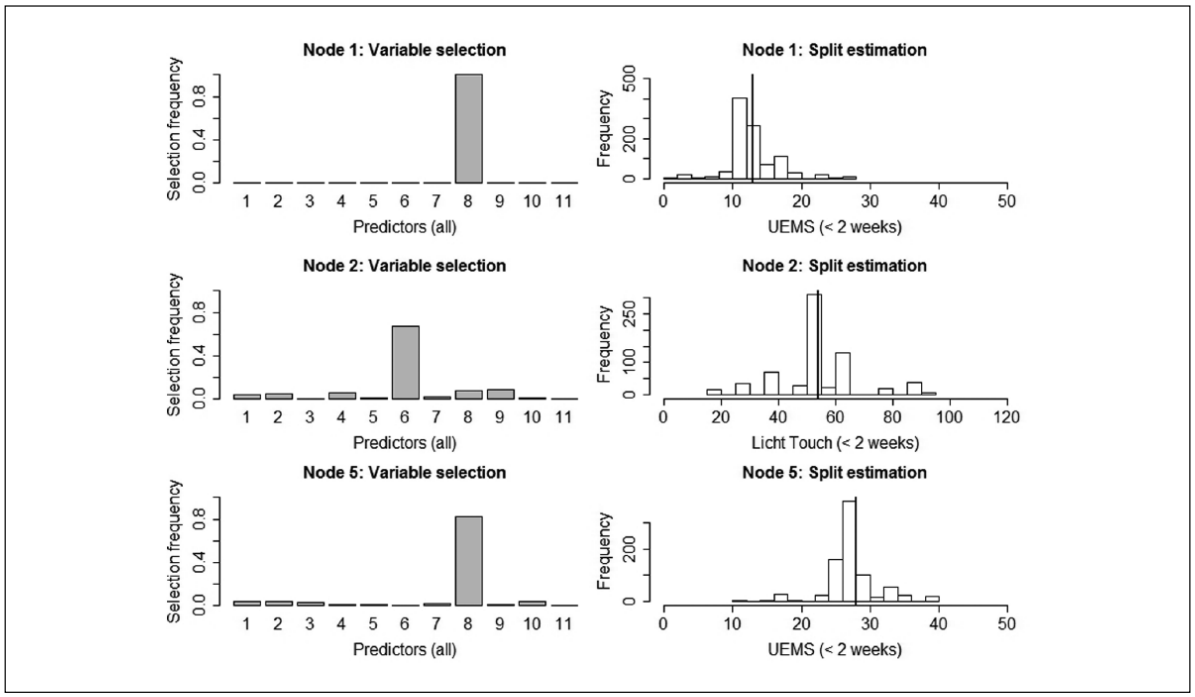

Figure 3. Internal validation of variable selection and split point estimation based on resampling.
For each inner node, predictor (early participant characteristic) selection (left panels) and split point estimation (right panels) are shown. The vertical line in the split point distribution represents the split implemented in the EMSCI-based conditional inference tree.
Abbreviations: EMSCI, European Multicenter study about Spinal Cord Injury; UEMS, upper extremity motor score.

For inner node 5, baseline UEMS was consistently selected, and no pattern of alternative variable selection emerged (predictor 8, bottom panel left, Figure 3). For all iterations where UEMS was the variable selected, the third split point distribution showed an almost symmetric distribution around the split point implemented in the EMSCI-based conditional inference tree (vertical line, bottom panel right).

External Validation

True validation of a statistical model can only be obtained by assessing its performance on an independent data set. We analyzed n = 83 participants from the Sygen[20] trial (from both control and treatment groups, as no significant difference was reported for these 2 different cohorts) by applying the decision rules provided by the inner nodes of the EMSCI-based tree and compared its terminal nodes to the results obtained for Sygen participants (Figure 4). Overall, we observed similar distributions for UEMS within terminal nodes for EMSCI and Sygen subjects.

Terminal node 6 was very similar for both data sets, whereas terminal nodes 3 and 7 showed a nonsignificant shift in distribution toward slightly higher UEMS values for Sygen participants. Terminal node 4 is similar for both data sets, especially in terms of interquartile range, but Sygen participants show a nonsignificant shift toward lower UEMS values. Terminal nodes 4 and 7 for Sygen as well as terminal node 3 for EMSCI represent a relatively small sample size.

Based on an asymptotic Wilcoxon Mann–Whitney rank sum test, a 95% confidence interval (CI) for the median difference (EMSCI − Sygen) was computed for each pair of terminal nodes. The sample estimate for difference in medians (hereafter med.diff) and relative 95% CIs were the following: med.diff (node 3) = −6.0 (CI: −13.0, 2.0); med.diff (node 4) = 3.0 (CI: −7.0, 13.0); med.diff (node 6) = 0.0 (CI: −4.0, 4.0); med.diff (node 7) = −2.0 (CI: −5.0, 0.0).

Discussion

The intent of this study was to determine whether well-informed and predictable stratification algorithms can be achieved for an acute or subacute SCI clinical trial. Previously, this has been difficult because of the highly variable spontaneous recovery patterns after iSCI. There is a strong need for a clear a priori stratification algorithm and, as seen with the current example, URP-CTREE may be useful in this regard. The advantage of this tree-based regression model is its ability to automatically stratify participants into relatively homogeneous subgroups on the basis of early clinical characteristics.

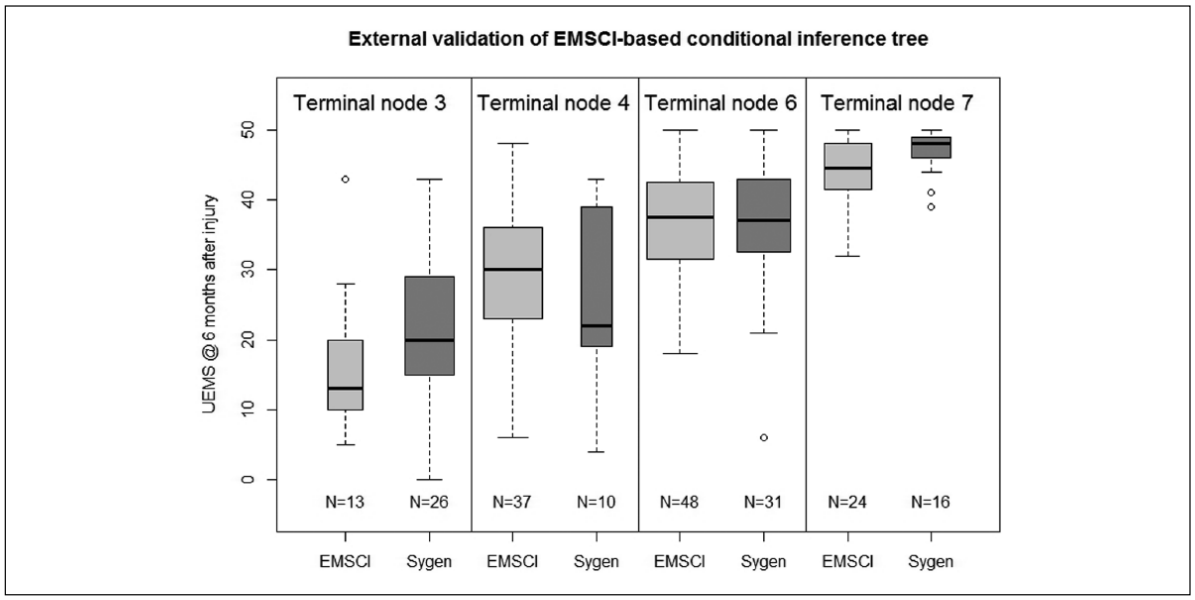

Figure 4. External validation of EMSCI-trained conditional inference tree based on Sygen participants with corresponding baseline characteristics.
EMSCI-derived terminal nodes (light gray) were compared to Sygen-derived terminal nodes (dark gray) provided by applying the splits provided by the EMSCI-based conditional inference tree. Boxplot width is proportional to available sample size, which is reported at the bottom.
Abbreviations: EMSCI, European Multicenter study about Spinal Cord Injury; UEMS, upper extremity motor score.

So far, very few studies in SCI have attempted to provide models for long-term prediction of neurological endpoints that could be used for patient stratification. To our knowledge, existing clinical algorithms refer to ambulatory[24,25] or functional[26] endpoints, which require substantial recovery and are therefore likely to depend on either a large treatment effect and/or spontaneous recovery. Thus, they are potentially less useful for application to an acute SCI clinical trial investigating a subtle therapeutic effect that would be initially detectable by some small neurological change. In addition, those functional endpoint studies are based on the statistical techniques of multiple linear and logistic regression. In a recent study,[19] we reported that those techniques do not provide direct rules for identifying more homogeneous cohorts and do not consider the full endpoint distribution. Here, we demonstrate the value of conditional inference trees[6] to provide an easily implementable, data-driven stratification rationale for the inclusion of iSCI into future SCI trial programs.

Conditional Inference Tree for EMSCI Data

The conditional inference tree for tetraplegic AIS-B and AIS-C based on EMSCI data produces 3 split points based on baseline UEMS and baseline light touch sensation (Figure 1). All 3 split points (inner nodes 1, 2, 5; Figure 1) are based on the most significant predictor (early participant characteristic) at that inner node.

As expected, baseline UEMS after iSCI plays an important role in predicting UEMS values 6 months later. A nonlinear relationship between baseline UEMS and the 6-month value is clearly shown by the 2 subsequent splits of baseline UEMS, as has been shown for general motor recovery in previous studies.[4,20] The lower range of UEMS is further divided depending on the perception of light touch. In our analysis, light touch was clearly more predictive for future UEMS than pin prick sensation (see Table 2). This seems to be in contrast with previous findings.[4,27] Those findings predicted ambulation for motor complete, sensory incomplete (AIS-B) participants based on their lower extremity and sacral pinprick score. Our results are therefore complementary rather than contrasting.

Although not always statistically significant, Table 2 also shows that the baseline motor level is consistently ranked within the first 3 most important predictors for each node. The relevance of the motor level for spontaneous motor recovery is dependent on the nature of SCI and has been reported before.[28] Its explicit incorporation in future analysis of motor endpoints is warranted.

AIS Grades for SCI Patient Stratification

Over the past few decades, stratification procedures for SCI trials were narrow,[2] broad,[3] or omitted.[14] Clinicians have been classifying SCI in wide-ranging categories, the most

common of which is the AIS classification. As can be seen in Table 1, the initial classification of iSCI severity as AIS-B or AIS-C had little predictive value for the distribution of participants with regard to their scores at 6 months. AIS-B and AIS-C participants were distributed across all cohorts, often fairly equally (Figure 1, nodes 3, 4, 6, 7). AIS grades have been repeatedly proven not to be a valuable measure for endpoint definition,[2,13,20] especially so in the context of SCI clinical trials with expected small treatment effects. Our analysis, even though based on AIS-B and AIS-C participants only, and other studies[29] show that AIS grades might not be appropriate for a fine-grained stratification as required by clinical trial either.

Validation

The main goal of any clinical predictive model is to provide valid endpoint predictions for future participants.[30] With the goal to investigate the stability and generalizability of conditional inference tree, we undertook internal and external validation of our results.

Internal validation was based on resampling and showed that predictor selection was consistent for all inner nodes, where the predictors selected in the EMSCI-based URP-CTREE were reliably selected (Figure 3, left panels) across resampling iterations. In addition, no pattern of selection for other predictors emerged from the resampling technique. The split point distribution for the predictors showed an approximately symmetric distribution around the original split (vertical line in right panels, Figure 3). Both predictor selection and split point distribution provide strong evidence for the stability of the EMSCI-based conditional inference tree.

To externally validate and support the general applicability of our conditional inference tree for UEMS, we applied the decision rules provided by the inner nodes of the EMSCI-based tree to an independent sample of Sygen participants (n = 83; from both control and treatment groups, as no significant difference was reported). Overall, we observed similar terminal nodes (final subgroups, Figure 4) for EMSCI and Sygen subjects, which provides evidence for the external validity of the EMSCI conditional inference tree. Similarities between European and North American spontaneous recovery profiles have already been reported for motor recovery after sensorimotor complete cervical (AIS-A) SCI.[31] Our analyses suggest that parallels extend to the iSCI population.

Specifically, nodes 3 and 4 (Figure 4) for Sygen patients provided a less sharp distinction between low (EMSCI boxplot node 3, Figure 4) and intermediate (EMSCI boxplot node 4, Figure 4) UEMS status at 6 months after injury. Terminal node 3 in Sygen showed higher UEMS scores than the corresponding EMSCI terminal node, while terminal node 4 presents the opposite behavior.

The different enrollment time frames of Sygen (first assessment within 3 days of SCI) compared to EMSCI (first assessment within 2 weeks, in average 8 days after injury) may contribute to any observed differences in terminal nodes. To support our hypothesis, we selected those EMSCI participants (n = 25) having an initial baseline assessment within 3 days of injury (ie, same time window as for Sygen), and we noted a median UEMS of 38 at 6 months. For those EMSCI participants that had an initial assessment >3 days after injury (n = 97), the median UEMS was 34 at 6 months. This analysis performed on the EMSCI data support our interpretation that enrollment time frame can influence UEMS outcomes.

Despite the observed differences (Figure 4), all 95% confidence intervals for the median difference of each terminal node pair included the value of zero, providing no evidence for a statistically significant shift in distribution in our data. Nonetheless, absence of evidence cannot be interpreted as evidence of absence, especially because of the small sample sizes.

While assessing external validity, we explicitly did not fit a new URP-CTREE to the Sygen data set as the goal of external validation is to see whether a fitted model provides valid prediction of an endpoint for an independent data set, and not whether the same model results from the 2 different datasets.

Applications to Clinical Trial Design

Our analysis was intended to provide an illustrative example of how URP-CTREE may be employed in the planning and designing of future clinical trials. Of the 4 nodes (3, 4, 6, 7) in Figure 1 and Table 1, a conservative enrollment strategy of iSCI participants in a phase II SCI clinical trial would only include participants within nodes 4 and 6. In this way, enrollment is restricted to iSCI participants that are likely to respond to a treatment, and participants where any treatment effect can be readily detected. In short, the assessment of a treatment effect is not hampered by floor or ceiling effects for the endpoint measurement. Participants in node 7 would likely be excluded because they will be constrained by a ceiling effect for the UEMS scale. For node 7, the median UEMS at 6 months is 44.5/50. Thus, only 5.5 motor points are available to detect a therapeutic effect in comparison to controls.

Participants in terminal node 3 could be excluded as well, because they show limited final upper extremity scores and any treatment targeting the enhancement of their spared sensorimotor function might only have a small effect. In addition, those participants are distinctly different from the other iSCI nodes in terms of final UEMS. In terms of UEMS and improvement in motor levels, the patterns for node 3 participants were very similar to those of cervical sensorimotor complete (AIS-A) participants and a similar

clinical endpoint (2 motor level improvement) might be justified.[9,32] Therefore, iSCI participants like those in terminal node 3 could be stratified with cervical AIS-A participants in a clinical trial with an appropriate endpoint.

The potential number of eligible iSCI participants that may be recruited to any clinical study is an important consideration. Even if we applied the more restrictive criteria and suggested only recruiting participants within nodes 4 and 6 (Figure 1), a hypothetical trial could potentially still recruit 70% of eligible iSCI subjects (EMSCI sample = 37 + 48/122 = 70%). Undoubtedly, a number of other inclusion and exclusion criteria will influence the final number of participants enrolled (the "funnel effect"[33]) but the potential inclusion of such a large percentage of iSCI combined with the a priori identification of more homogeneous subgroups represents a clear improvement in trial design.

Modeling Specific Endpoints

Independent from its use in clinical trials, one of the key advantages of URP-CTREE with respect to established statistical approaches[24-26] is that it presents a complete endpoint distribution within the terminal nodes.[19] This allows the exploration of different neurological or functional endpoints and their appropriateness for distinct iSCI subgroups (terminal nodes).

Clearly, the use of URP-CTREE does not guarantee partitioning of a heterogeneous population in the same way as presented here. URP-CTREE is specific for a given patient population, clinical endpoint, and early clinical predictors (baseline variables). Changes in any of these components may influence the resulting partitioning by generating different splits, a different number of terminal nodes (subgroups), or a different endpoint distribution within each node.

With any comprehensive historical database, URP-CTREE can be employed to accomplish all the necessary modeling variations and gather insight into the expected behavior of the control group for any chosen endpoint. This is a major improvement and should provide additional confidence for the final trial protocol submission. In the study reported here, we have deliberately not chosen a threshold value for UEMS at 6 months after iSCI. A threshold for a binary endpoint could be explored using URP-CTREE, but selection of any endpoint threshold requires consideration of what is reasonable and/or meaningful.

Limitations

URP-CTREE is applicable to several types of regression problems, including nominal, ordinal, numeric, and censored endpoints.[6] It remains to be demonstrated whether neurological scores like UEMS can actually be analyzed as a continuous endpoint, such as time or distance. Despite being the current default approach in SCI, total UEMS is in fact a sum of several ordinal variables (each manual muscle score, from a key muscle, is an ordinal scale in its own right). In other fields, it has been shown that the analysis of an ordinal scale with methods designed for the analysis of continuous endpoints led to errors in inference.[34-36] We are analyzing methods for correcting this limitation.

Conclusion

Our analyses provide an example for the employment of URP-CTREE to facilitate the design of more inclusive clinical trials. We utilized a conditional inference tree to provide an early prediction-based stratification of cervical iSCI patients as anchored by UEMS at 6 months after injury (inclusion/exclusion criteria). Internal validation of the EMSCI determined URP-CTREE proved it to be very stable. External validation based on comparable incomplete SCI participants from the independent Sygen study supports its generalizability. Our analysis supports the use of URP-CTREE as a statistical tool for modeling various other clinical endpoints and different patient populations. Further investigations on the external validity and the analysis of ordinal variables with techniques designed for continuous endpoint should be the object of refined analysis.

Acknowledgments

We appreciate the continuous assistance of René Koller with the EMSCI database.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors acknowledge the support of the European Multicenter study about Spinal Cord Injury (EMSCI), the International Foundation for Research in Paraplegia (IFP), the Spinal Cord Outcomes Partnership Endeavor (SCOPE), and the Clinical Research Priority Program in Neuro-rehabilitation (CRPP NeuroRehab) of the University of Zurich.

References

1. Wyndaele M, Wyndaele JJ. Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey? *Spinal Cord*. 2006;44:523-529.
2. Lammertse DP, Jones LAT, Charlifue SB, et al. Autologous incubated macrophage therapy in acute, complete spinal cord injury: results of the phase 2 randomized controlled multicenter trial. *Spinal Cord*. 2012;50:661-671.

3. Casha S, Zygun D, McGowan MD, Bains I, Yong VW, John Hurlbert R. Results of a phase II placebo-controlled randomized trial of minocycline in acute spinal cord injury. *Brain*. 2012;135:1224-1236.
4. Fawcett JW, Curt A, Steeves JD, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. *Spinal Cord*. 2006;45:190-205.
5. Yu KD, Zhu R, Zhan M, et al. Identification of prognosis-relevant subgroups in patients with chemoresistant triple-negative breast cancer. *Clin Cancer Res*. 2013;19: 2723-2733.
6. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat*. 2006;15. http://amstat.tandfonline.com/doi/full/10.1198/106186006X133933. Accessed January 20, 2014.
7. Kirshblum SC, Waring W, Biering-Sorensen F, et al. Reference for the 2011 revision of the international standards for neurological classification of spinal cord injury. *J Spinal Cord*. 2011;34:547-554.
8. Pouw MH, van Middendorp JJ, van Kampen A, et al. Diagnostic criteria of traumatic central cord syndrome. Part 1: A systematic review of clinical descriptors and scores. *Spinal Cord*. 2010;48:652-656.
9. Kramer JLK, Lammertse DP, Schubert M, Curt A, Steeves JD. Relationship between motor recovery and independence after sensorimotor-complete cervical spinal cord injury. *Neurorehabil Neural Repair*. 2012;26:1064-1071.
10. Al-Habib AF, Attabib N, Ball J, Bajammal S, Casha S, Hurlbert RJ. Clinical predictors of recovery after blunt spinal cord trauma: systematic review. *J Neurotrauma*. 2011;28:1431-1443.
11. Wilson JR, Cadotte DW, Fehlings MG. Clinical predictors of neurological outcome, functional status, and survival after traumatic spinal cord injury: a systematic review. *J Neurosurg Spine*. 2012;17:11-26.
12. Rudhe C, van Hedel HJA. Upper extremity function in persons with tetraplegia: relationships between strength, capacity, and the spinal cord independence measure. *Neurorehabil Neural Repair*. 2009;23:413-421.
13. Steeves JD, Lammertse D, Curt A, et al. Guidelines for the conduct of clinical trials for spinal cord injury (SCI) as developed by the ICCP panel: clinical trial outcome measures. *Spinal Cord*. 2006;45:206-221.
14. Bracken MB, Shepard MJ, Collins WF, et al. A randomized, controlled trial of methylprednisolone or naloxone in the treatment of acute spinal-cord injury—results of the Second National Acute Spinal Cord Injury Study. *N Engl J Med*. 1990;322:1405-1411.
15. Costello TJ, Swartz MD, Sabripour M, Gu X, Sharma R, Etzel CJ. Use of tree-based models to identify subgroups and increase power to detect linkage to cardiovascular disease traits. *BMC Genet*. 2003;4:S66.
16. Owzar K. Alternate statistical tools and limitations in genetic marker association studies in single-arm drug cancer trials. *J Clin Oncol*. 2008;26:1400-1401.
17. De Roock W, Claes B, Bernasconi D, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol*. 2010;11:753-762.
18. Brown AW, Malec JF, McClelland RL, Diehl NN, Englander J, Cifu DX. Clinical elements that predict outcome after traumatic brain injury: a prospective multicenter recursive partitioning (decision-tree) analysis. *J Neurotrauma*. 2005;22:1040-1051.
19. Tanadini LG, Steeves JD, Hothorn T, et al. Identifying homogeneous subgroups in neurological disorders: unbiased recursive partitioning in cervical complete spinal cord injury. *Neurorehabil Neural Repair*. 2014;28:507-515.
20. Geisler FH, Coleman WP, Grieco G, Poonian D, Group SS. The SYGEN® multicenter acute spinal cord injury study. *Spine*. 2001;26:S87-S98.
21. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/. Accessed January 21, 2015.
22. Hothorn T, Hornik K, Strobl C, Zeileis A. party: a laboratory for recursive partitioning (R package, Version 1.0-13). http://cran.r-project.org/package=party. Accessed January 21, 2015.
23. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. coin: conditional inference procedures in a permutation test framework (R package, Version 1.0-23). http://cran.r-project.org/package=coin. Accessed January 21, 2015.
24. Zörner B, Blanckenhorn WU, Dietz V, Curt A. Clinical algorithm for improved prediction of ambulation and patient stratification after incomplete spinal cord injury. *J Neurotrauma*. 2010;27:241-252.
25. Van Middendorp JJ, Hosman AJ, Donders ART, et al. A clinical prediction rule for ambulation outcomes after traumatic spinal cord injury: a longitudinal cohort study. *Lancet*. 2011;377:1004-1010.
26. Wilson JR, Grossman RG, Frankowski RF, et al. A clinical prediction model for long-term functional outcome after traumatic spinal cord injury based on acute clinical and imaging factors. *J Neurotrauma*. 2012;29:2263-2271.
27. Oleson CV, Burns AS, Ditunno JF, Geisler FH, Coleman WP. Prognostic value of pinprick preservation in motor complete, sensory incomplete spinal cord injury. *Arch Phys Med Rehabil*. 2005;86:988-992.
28. Coleman W. Injury severity as primary predictor of outcome in acute spinal cord injury: retrospective results from a large multicenter clinical trial. *Spine J*. 2004;4:373-378.
29. Velstra IM, Bolliger M, Tanadini LG, et al. Prediction and stratification of upper limb function and self-care in acute cervical spinal cord injury with the graded redefined assessment of strength, sensibility, and prehension (GRASSP). *Neurorehabil Neural Repair*. 2014;28:632-642.
30. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (Auflage: Softcover reprint of hardcover 1st ed). New York, NY: Springer; 2009.

31. Steeves JD, Kramer JK, Fawcett JW, et al. Extent of spontaneous motor recovery after traumatic cervical sensorimotor complete spinal cord injury. *Spinal Cord*. 2011;49:257-265.
32. Steeves JD, Lammertse DP, Kramer JLK, et al. Outcome measures for acute/subacute cervical sensorimotor complete (AIS-A) spinal cord injury during a phase 2 clinical trial. *Top Spinal Cord Inj Rehabil*. 2012;18:1-14.
33. Jones LAT, Lammertse DP, Charlifue SB, et al. A phase 2 autologous cellular therapy trial in patients with acute, complete spinal cord injury: pragmatics, recruitment, and demographics. *Spinal Cord*. 2010;48:798-807.
34. Winship C, Mare RD. Regression models with ordinal variables. *Am Sociol Rev*. 1984;512-525.
35. Hastie TJ, Botha JL, Schnitzler CM. Regression with an ordered categorical response. *Stat Med*. 1989;8:785-794.
36. Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. *J Clin Epidemiol*. 1997;50:45-55.

PAPER III

Autoregressive transitional ordinal model to test for treatment effect in neurological trials with complex endpoints

Lorenzo G. Tanadini, John D. Steeves, Armin Curt, Torsten Hothorn

Paper published in *BMC Medical Research Methodology*, 2016, 16-149.

Tanadini *et al. BMC Medical Research Methodology* (2016) 16:149
DOI 10.1186/s12874-016-0251-y

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Autoregressive transitional ordinal model to test for treatment effect in neurological trials with complex endpoints

Lorenzo G. Tanadini[1*], John D. Steeves[2], Armin Curt[3] and Torsten Hothorn[1]

Abstract

Background: A number of potential therapeutic approaches for neurological disorders have failed to provide convincing evidence of efficacy, prompting pharmaceutical and health companies to discontinue their involvement in drug development. Limitations in the statistical analysis of complex endpoints have very likely had a negative impact on the translational process.

Methods: We propose a transitional ordinal model with an autoregressive component to overcome previous limitations in the analysis of Upper Extremity Motor Scores, a relevant endpoint in the field of Spinal Cord Injury. Statistical power and clinical interpretation of estimated treatment effects of the proposed model were compared to routinely employed approaches in a large simulation study of two-arm randomized clinical trials. A revisitation of a key historical trial provides further comparison between the different analysis approaches.

Results: The proposed model outperformed all other approaches in virtually all simulation settings, achieving on average 14 % higher statistical power than the respective second-best performing approach (range: -1 %, +34 %). Only the transitional model allows treatment effect estimates to be interpreted as conditional odds ratios, providing clear interpretation and visualization.

Conclusion: The proposed model takes into account the complex ordinal nature of the endpoint under investigation and explicitly accounts for relevant prognostic factors such as lesion level and baseline information. Superior statistical power, combined with clear clinical interpretation of estimated treatment effects and widespread availability in commercial software, are strong arguments for clinicians and trial scientists to adopt, and further extend, the proposed approach.

Keywords: Upper extremity motor scores, Summed overall score, Multivariate ordinal endpoints, Proportional odds model, Statistical power, Spinal cord injury, Sygen®trial, Rasch models, Latent variable models

Background

Neurological research is responsible for the investigation of many devastating disorders such as stroke, Alzheimer's and Parkinson's diseases. In terms of health costs, brain-related disorders are a greater socio-economic burden than cancer, cardiovascular diseases and diabetes combined [1], with yearly costs for the European society estimated at almost 400 billion € [2].

Despite several therapeutic approaches [3–6] based on recent discoveries of cellular and molecular processes of degeneration, but also spontaneous regeneration following injury, pharmaceutical and health companies have been withdrawing from neuroscience, as a number of trials intended to show efficacy of treatments for neurological disorders failed [7]. In the field of Spinal Cord Injury (SCI), four decades after the first pharmacological treatment of acute injuries [8], the promises of preclinical discoveries have yet to be translated into a standard treatment [9].

To streamline the translational process, the International Campaign for Cures of Spinal Cord Injury Paralysis

*Correspondence: lorenzo.tanadini@uzh.ch
[1]Department of Biostatistics; Epidemiology, Biostatistics and Prevention Institute; University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland
Full list of author information is available at the end of the article

(ICCP) appointed in 2007 an international panel with the task to reviewing strengths and weaknesses of clinical trials in spinal cord injury. Their recommendations for the planning and conduction of future trials were condensed in a series of publications [10–13], which strongly influenced the conception of clinical trials thereafter [14].

Nonetheless, the ICCP reviews [10–13] did not solicit the application of the most appropriate and recent statistical techniques available for the analysis of complex SCI trial endpoints, and many clinical trials failed to do so too [15–19].

In fact, virtually all routinely performed clinical assessments in spinal cord injury are measured on ordinal scales, which are characterized by an arbitrary numerical score establishing a ranking of observations. The difference between two following ranks is by no means bound to be equivalent across the range of the scale, preventing standard operations such as addition, and making the use of statistical methods developed for continuous endpoints inappropriate. Despite this, clinical trials designed and powered for a primary ordinal endpoint often resorted to adding several ordinal endpoints to form a single overall summed score, which is in some cases subsequently collapsed to a binary outcome [15–19]. These approaches have been shown to be inappropriate in a number of aspects [20], and practical consequences such as biased parameter estimates, misleading associations and loss of power are some of the known consequences of assuming metric properties for ordinal endpoints [21–23].

In this study, we propose for the first time in SCI a transitional ordinal model with an autoregressive component for testing for treatment effect on a multivariate ordinal endpoint such as the Upper Extremity Motor Scores (UEMS), while comparing it to current analysis approaches in terms of statistical power and clinical interpretation of treatment effect estimates.

Methods

The objective was to propose a new approach to the analysis of complex ordinal endpoints in neurological clinical trials, and provide statistical power comparisons of procedures for treatment effect testing. Two-armed Randomized Clinical Trials (RCT) with specific levels of experimental conditions were generated and analysed. Current approaches to the analysis of multivariate ordinal endpoints such as the Upper Extremity Motor Scores (UEMS) were compared to the proposed autoregressive transitional ordinal model. The proposed approach models the transition, e.g. the change in UEMS distribution, from trial baseline to trial end. The autoregressive term of the model describes the anatomical structure of the spinal cord by postulating a direct dependency between contiguous segments.

Data source and trial endpoint

The data utilized in this study was extracted from the European Multicenter Study about Spinal Cord Injury (EMSCI, ClinicalTrials.gov Identifier: NCT01571531, www.emsci.org). EMSCI tracks the functional and neurological recovery of patients during the first year after spinal cord injury in a highly standardized manner. All patients gave written informed consent. The ethical committee of the Canton of Zurich, Switzerland, has previously approved the EMSCI project, upon which this project is based, and the approval is also valid for any statistical analysis/re-analysis.

To reflect the time frame of a possible future clinical trial, we considered baseline (within 2 weeks after injury, $t = 1$) and one follow-up (6 months after injury, $t = 2$) examination. For this simulation study, we extracted and utilized records of N=405 patients with a Motor Level (ML) defined between spinal segments C5-T1 (see Additional file 1 for details) and with available baseline information.

The trial endpoint considered is the Upper Extremity Motor Scores. UEMS represents a subset of the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI) [24] and describes the muscle contraction force for 10 key muscles on the arms and hands (5 on each body side), each one being rated on a 6-point ordinal scale (0: total paralysis, through 5: active movement against full resistance, see Additional file 1 for details). Accordingly, $Y_{i,m,t}$ is the muscle contraction score for patient i ($i = 1, \ldots, n$) and key muscle m ($m = 1, \ldots, 10$) measured at time point t ($t = 1, 2$). Each key muscle $Y_{i,m,t}$ is therefore an ordinal variable with $k = 6$ levels $0 < 1 < \ldots < 5$, and UEMS is a multivariate ordinal endpoint. The chosen endpoint is particularly relevant in SCI. A change in total UEMS over trial period has been employed repeatedly in clinical trials [15, 19] and has been suggested to correlate with changes in activities of daily living that rely on recovery of upper extremity function [25].

RCT simulation

An autoregressive transitional ordinal model of the form:

$$logit\left[P(y_{i,m,2} \leq k)\right] = \alpha_j + \beta_{\text{lev}}\, x_{\text{lev},i,m,1} + \beta_{\text{base}}\, y_{\text{base},i,m,1} + \beta_{\text{auto}}\, y_{\text{auto},i,m-1,2} \quad (1)$$

was fitted on the EMSCI data. α_j are the $k - 1 = 5$ intercept parameters, x_{lev} is a 10-level nominal factor denoting the combination of Motor Level and the distance from the Motor Level to the key muscle m being analysed, expressed as number of key muscles along the spine (reference: motor level: cervical C5, distance: -1 (first muscle

below the level)), $y_{\text{base},i,m,1}$ is the ordered factor for baseline motor score of key muscle m, and $y_{\text{auto},i,m-1,2}$ is the ordered factor for motor score of the key muscle just above the one being analysed at $t = 2$. The autoregressive term of the model describes the anatomical structure of the spinal cord, and postulates that the motor score of a given key muscle depends on the Motor Score of the key muscle just rostral to it. As a consequence, the observed pattern of lower motor scores with increasing distance from the ML is reproduced. In accordance with the above description, Eq. 1 simulated and analysed only key muscle score below the Motor Level. Motor scores $y_{i,m,2}$ for key muscles at ML were multinomially sampled from corresponding observed EMSCI frequencies at Motor Level, while motor scores $y_{i,m,2}$ for key muscles above the ML were given the maximal score.

The parameter estimates recovered from the model specified in Eq. 1 describe the spontaneous neurological recovery for patients under standard of care and were subsequently used to simulate participants in the control arm of the trial. From the EMSCI data we also computed the observed frequencies of Motor Level combinations for the left and right body side at baseline. Given that patients having both left and right ML at the lowest UEMS key muscles T1 are very rare (3 % in our EMSCI sample) and do not contribute to the analysis (no key muscles in the UEMS below the ML), they were not included into the simulation.

Equation 1 models the spontaneous neurological recovery for patients under standard of care. We introduced an additional parameter β_{trt} representing a postulated treatment effect, leading to an autoregressive transitional ordinal model of the form:

$$logit\left[P(y_{i,m,2} \leq k)\right] = \alpha_j + \beta_{\text{lev}}\, x_{\text{lev},i,m,1} + \beta_{\text{base}}\, y_{\text{base},i,m,1} + \beta_{\text{auto}}\, y_{\text{auto},i,m-1,2} + \beta_{\text{trt}}\, x_{\text{trt},i,1} \quad (2)$$

As previously defined, α_j are the $k - 1 = 5$ intercept parameters, x_{lev} is a 10-level nominal factor denoting the combination of Motor Level and the distance from the Motor Level to the key muscle m being analysed, expressed as number of key muscles along the spine (reference: Motor Level: C5, distance: -1), $y_{\text{base},i,m,1}$ is the ordered factor for baseline motor score of key muscle m, $y_{\text{auto},i,m-1,2}$ is ordered factor for motor score of the key muscle just above the one being analysed at $t = 2$, and x_{trt} is an indicator for treatment arm with placebo as reference.

The autoregressive term of the model describes the anatomical structure of the spinal cord, and postulates that the motor score of a given key muscle depends on the motor score of the key muscle just rostral to it. As a consequence, the observed pattern of lower motor scores with increasing distance from the ML is reproduced. Besides the postulated treatment effect β_{trt}, which is set to different values depending on the simulation settings, all other parameters in Eq. 2 were kept equal to the estimates recovered by fitting Eq. 1 to the EMSCI data.

We thus simulated randomized clinical trials with two treatment arms and specific levels of experimental conditions. To cover possible SCI early phase as well as phase III settings, we generated total trial sample sizes of 50, 75, 100, 125, 150, 175, 200 participants. To our knowledge, there is to date no publication on the magnitude of possible treatment effects for UEMS which could have guided us in defining more tailored scenarios. We therefore postulated a rather wide range of six possible treatment effects (from no treatment effect ($\beta_{\text{trt}} = 0.0 = \log(1)$) to strong treatment effect ($\beta_{\text{trt}} = 0.4055 = \log(1.5)$) in 0.1 steps). A total of 42 scenarios resulted from simulating all possible combinations of the 7 trial sample sizes and 6 possible treatment effects considered. Being a proportional odds model, the exponentiated β_{trt} can be interpreted as conditional Odds Ratio (OR) between trial arms, meaning that, conditional on all other prognostic factors being equal, it specifies the ratio of the odds for a key muscle to achieve a motor score of less than or equal to k in the treatment arm divided by the same odds in the control arm. OR is a statistically sensible and clinical widely accepted way of quantifying effects of categorical variables.

The 42 trial scenarios resulting from all combinations of 7 trial sample sizes and 6 possible treatment effects were simulated in the following way:

1. Right and left Motor Levels for the hypothesized number of trial participants were drawn from a multinomial distribution with category probabilities set to the corresponding observed EMSCI frequencies.
2. Baseline UEMS for each trial participant were sampled with replacement from all EMSCI patients having the same left-right ML constellation.
3. Each simulated participant was randomly allocated to either the control or the treatment arm with a 1:1 allocation scheme.
4. UEMS at six months for the key muscle at ML were drawn from a multinomial distribution with category probabilities set to the corresponding observed EMSCI frequencies.
5. UEMS at six months below the ML were simulated using the previously fitted model for spontaneous recovery (Eq. 1) for participants in the control arm, and the same model with the addition of a postulated treatment effect (Eq. 2) for participants in the treatment arm of the trial.

6. Each one of the 42 trial scenarios was replicated 1000 times.
7. A battery of 6 different tests for treatment effect (see below "Endpoint analysis approaches" Section) were applied to each simulated trial.
8. The statistical power = P(reject $H_0|H_1$ is true) was estimated as the fraction of significant tests for treatment effect at the nominal level 0.05 among the 1000 replications.

Endpoint analysis approaches

In neurology in general, and SCI in particular, very common approaches to the analysis of UEMS or similar endpoints are as the total sum of all motor scores $Y^*_{i,2} = \sum_{m=1}^{10} Y_{i,m,2}$ or as difference between two time points $Y^{**}_i = \sum_{m=1}^{10} Y_{i,m,2} - Y_{i,m,1}$. Accordingly, treatment effect for UEMS was tested with:

t-test: t-test for $Y^*_{i,2}$, comparing mean total UEMS in the two treatment groups.

t-test delta: t-test for Y^{**}_i, comparing the mean difference in total UEMS from baseline to the end of the trial between the two treatment groups.

ANCOVA: Analysis of covariance for $Y^*_{i,2}$, comparing mean total UEMS in the two treatment groups with baseline total UEMS $Y^*_{i,1}$ as controlling continuous variable.

Even though not commonly done in SCI, we considered necessary that the Motor Level should be incorporated into the analysis of motor function. In fact, its importance has been reported before [26, 27]. We therefore applied a conditional test of independence between outcome and treatment arm which was stratified according to the Motor Level of each trial participant. We predicted that this approach would perform better than the previous, not stratified ones, and explored the possibility to utilise them as "ad hoc" approach for the analysis of UEMS. Accordingly, treatment effect for UEMS was tested with:

i-test: stratified independence test for $Y^*_{i,2}$, comparing total UEMS in the two treatment groups.

i-test delta: stratified independence test for Y^{**}_i, comparing the difference in total UEMS from baseline to the end of the trial between the two treatment groups.

Both tests are implemented in the R add-on package **coin** [28, 29].

The last approach for the analysis of UEMS in a RCT is a model that takes into account the ordinal nature of each key muscle and explicitly incorporates baseline UEMS as well as ML into the analysis:

transitional: transitional ordinal model for $Y_{i,m,2}$ of the form specified in Eq. 2, comparing the shift in motor score probabilities associated with treatment.

The proposed model is a proportional odds model with an autoregressive component. The latter takes into account the spatial orientation of the key muscles along the spinal cord by postulating a direct dependency of adjacent spinal segments. As a consequence, the observed pattern of lower Motor Scores with increasing distance from the ML is reproduced. This model was fitted using function `polr` from the R add-on package **MASS** [30, 31].

The parameter β_{trt}, which quantifies the treatment effect on the link scale, is the focus of the proposed model. Its significance testing was based on a permutation test [32, 33], where the distribution of the test statistics under H_0 (no treatment effect) was based on refitting the same model 1000 times after randomly rearranging the labels for arm allocation. This type of statistical significance test does not rely on any distributional assumption. In addition, by permuting trial arm allocation at participant level, we accounted for the hierarchical structure of the data analysed, where multiple key muscles are measured on the same participant. All computations were performed in the R system for statistical computing [34], version 3.1.3. The R code implementing the simulation study is available online (doi: http://dx.doi.org/10.5281/zenodo.47600).

Revisiting a key SCI trial

As a practical application, we analysed a subset of the data collected during a past clinical trial. The Sygen ®trial recruited N=760 SCI participants in 28 centres in North-America in a 5-year period between 1992 and 1997 [17, 35, 36]. Sygen ®is a naturally occurring compound in cell membranes which has been associated with neuroprotective and regenerative effects in a number of experimental models and early-phase human trials. The trial is an example where a promising therapeutic approach was finally abandoned, as no significant treatment effect could be assessed on the primary endpoint despite a considerable final sample size (N=760). The primary endpoint assessed the overall neurological status of a patient and was defined as a dichotomization derived from an ordinal scale (see [36] for the exact definition). The primary endpoint was analysed by means of logistic regression. Several ancillary analyses were performed and mostly preferred the treatment arm, even though the differences were not always statistically significant. To our knowledge, no analysis performed at the level of motor scores of the upper extremity key muscles UEMS as reported here have been published.

We revisited the trial by testing for treatment effect on the UEMS with all six approaches outlined before (see "Endpoint analysis approaches" Section). The proposed

autoregressive transitional ordinal model (Eq. 2) can be easily fitted as proportional odds model to the segment-wise UEMS data in the long format. The autoregressive component $y_{i,m-1,2}$ can be incorporated by shifting the six-month, muscle-wise UEMS entries so as to be aligned to the key muscle $y_{i,m,2}$ just caudal to them.

To reflect our simulation study, we selected participants with a ML between C5 - C8 (T1 were discarded, because there is no key muscle caudal to the ML on the UEMS), and considered only patients treated with a low dosage (the original trial had two treatment doses, the higher of which was abandoned during the study). After patients selection, we analysed a finale sample of N=284 participants, 127 (45 %) of which in the control arm. This analysis is intended to give an example of the application of the proposed transitional ordinal model, but is not intended and should not be taken as a definitive conclusion about the value or outcome of the trial. Given the strongly selected patients sample utilised, the different endpoint analysed and the different scope of our analysis, generalizations of this type cannot be drawn.

Results

RCT simulation

For the purpose of this study, we simulated 1000 times each one of the 42 different combinations of trials size and postulated treatment effect. Statistical power, which is defined as the probability of rejecting the H_0 of no treatment when there is in fact a treatment effect, was estimated as the fraction of this 1000 iterations where the test for treatment effect resulted significant at the 0.05 level. Table 1 reports the statistical power of all treatment testing approaches for all simulated settings. Figure 1 shows the statistical power of all six approaches for the intermediate treatment effect simulated. Figure 2 displays graphically the statistical power of all treatment testing approaches for all simulated settings. The nominal level 0.05 was maintained by all approaches when no treatment effect was introduced in the simulation, making further comparisons between different approaches straightforward.

For the smallest treatment effect $\beta_{\text{trt}} = 0.0953 = \log(1.1)$, all six tests for treatment effect showed a low power, never exceeding $P(\text{reject } H_0|H_1 \text{ is true}) \leq 0.135$. The transitional ordinal model was nonetheless superior to all other approaches in virtually every trial size setting, its power point estimates averaging 2.3 % higher than the respective second best-performing approach.

Already at the next higher treatment effect simulated $\beta_{\text{trt}} = 0.1823 = \log(1.2)$, the transitional ordinal model showed roughly twice as much power as the second-best performing approach, though it did not exceed $P(\text{reject } H_0|H_1 \text{ is true}) \leq 0.36$. This held for all simulation settings except the smallest sample size. Statistical power point estimates for the transitional ordinal model were on average 10.3 % higher than the respective second best-performing approach.

In the settings with median simulated treatment effect $\beta_{\text{trt}} = 0.2624 = \log(1.3)$ shown in Fig. 1, the transitional ordinal model was superior for all trial sizes. Power point estimates for the proposed model were on average 19.4 % higher than the respective second best-performing approach, with this difference in performance increasing with increasing trial size.

With the simulated treatment effect of $\beta_{\text{trt}} = 0.3365 = \log(1.4)$, the transitional ordinal model had superior statistical power of 26.3 % on average, compared to the respective second best-performing approach, with this difference increasing with increasing trial size.

For the largest simulated treatment effect of $\beta_{\text{trt}} = 0.4055 = \log(1.5)$, the transitional ordinal model had an average superior statistical power of 27.9 %, compared to the respective second best-performing approach. The difference in performance increased strongly up to trial size N=100, but then declined with larger sizes.

Overall, despite a comparably poor performance of all approaches for small simulated treatment effects, a stable pattern in the ranking of performance emerged: the proposed transitional ordinal approach provided best power results in virtually all settings. ANCOVA was usually the second-best approach, closely followed by the independence test on the difference of UEMS from baseline Y_i^{**}, the similarly performing t-test on the difference of UEMS from baseline Y_i^{**} and the independence test on the UEMS after six months $Y_{i,2}^{*}$. The t-test on the UEMS after six months $Y_{i,2}^{*}$ performed worst in almost all settings.

Revisiting a key SCI trial

We analysed a subset of the data collected during the Sygen ®trial [17, 35, 36]. To our knowledge, no analysis on this data has been performed at the level of motor scores of the upper extremity key muscles UEMS as reported here. The results of the six analysis approaches (see Endpoint analysis approaches section) are reported here:

t-test: No significant difference in the estimated means $\widehat{\mu_{\text{ctrl}}} = 30.370$ and $\widehat{\mu_{\text{trt}}} = 30.170$ of UEMS at 6 months between trial arms: t(275)=0.130, $p-$value = 0.896.

t-test delta: No significant difference in the estimated mean change $\widehat{\mu_{\text{ctrl}}} = 11.978$ and $\widehat{\mu_{\text{trt}}} = 10.540$ of UEMS between trial arms: t(259)=1.239, $p-$value = 0.216.

ANCOVA: No significant difference in the estimated means of UEMS at 6 months between trial arms,

Table 1 Statistical power for all simulation settings. Point estimates, as well as Wilson confidence intervals are reported for all analysis approaches

Size	Treatment	OR	T-test	CI lower	CI upper	T-test delta	CI lower	CI upper	I-test	CI lower	CI upper	I-test delta	CI lower	CI upper	ANCOVA	CI lower	CI upper	Transitional	CI lower	CI upper
50	0.0000	1.0	0.053	0.041	0.069	0.052	0.040	0.068	0.051	0.039	0.066	0.042	0.031	0.056	0.046	0.035	0.061	0.050	0.038	0.065
75	0.0000	1.0	0.048	0.036	0.063	0.050	0.038	0.065	0.052	0.040	0.068	0.051	0.039	0.066	0.053	0.041	0.069	0.052	0.040	0.068
100	0.0000	1.0	0.047	0.036	0.062	0.046	0.035	0.061	0.054	0.042	0.070	0.046	0.035	0.061	0.048	0.036	0.063	0.045	0.034	0.060
125	0.0000	1.0	0.049	0.037	0.064	0.052	0.040	0.068	0.040	0.030	0.054	0.056	0.043	0.072	0.056	0.043	0.072	0.057	0.044	0.073
150	0.0000	1.0	0.056	0.043	0.072	0.044	0.033	0.059	0.041	0.030	0.055	0.040	0.030	0.054	0.050	0.038	0.065	0.040	0.030	0.054
175	0.0000	1.0	0.050	0.038	0.065	0.050	0.038	0.065	0.043	0.032	0.057	0.053	0.041	0.069	0.042	0.031	0.056	0.047	0.036	0.062
200	0.0000	1.0	0.051	0.039	0.066	0.052	0.040	0.068	0.046	0.035	0.061	0.053	0.041	0.069	0.056	0.043	0.072	0.048	0.036	0.063
50	0.0953	1.1	0.057	0.044	0.073	0.060	0.047	0.076	0.063	0.050	0.080	0.052	0.040	0.068	0.062	0.049	0.079	0.049	0.037	0.064
75	0.0953	1.1	0.055	0.042	0.071	0.056	0.043	0.072	0.051	0.039	0.066	0.069	0.055	0.086	0.049	0.037	0.064	0.086	0.070	0.105
100	0.0953	1.1	0.057	0.044	0.073	0.071	0.057	0.089	0.061	0.048	0.078	0.071	0.057	0.089	0.071	0.057	0.089	0.106	0.088	0.127
125	0.0953	1.1	0.074	0.059	0.092	0.068	0.054	0.085	0.082	0.067	0.101	0.075	0.060	0.093	0.081	0.066	0.100	0.094	0.077	0.114
150	0.0953	1.1	0.063	0.050	0.080	0.070	0.056	0.088	0.062	0.049	0.079	0.075	0.060	0.093	0.078	0.063	0.096	0.116	0.098	0.137
175	0.0953	1.1	0.066	0.052	0.083	0.071	0.057	0.089	0.069	0.055	0.086	0.079	0.064	0.097	0.073	0.058	0.091	0.117	0.099	0.138
200	0.0953	1.1	0.072	0.058	0.090	0.101	0.084	0.121	0.080	0.065	0.098	0.092	0.076	0.112	0.099	0.082	0.119	0.135	0.115	0.158
50	0.1823	1.2	0.068	0.054	0.085	0.090	0.074	0.109	0.065	0.051	0.082	0.091	0.075	0.110	0.093	0.077	0.113	0.111	0.093	0.132
75	0.1823	1.2	0.096	0.079	0.116	0.095	0.078	0.115	0.106	0.088	0.127	0.100	0.083	0.120	0.107	0.089	0.128	0.164	0.142	0.188
100	0.1823	1.2	0.106	0.088	0.127	0.098	0.081	0.118	0.112	0.094	0.133	0.099	0.082	0.119	0.114	0.096	0.135	0.226	0.201	0.253
125	0.1823	1.2	0.115	0.097	0.136	0.127	0.108	0.149	0.135	0.115	0.158	0.132	0.112	0.154	0.145	0.125	0.168	0.261	0.235	0.289
150	0.1823	1.2	0.134	0.114	0.157	0.155	0.134	0.179	0.138	0.118	0.161	0.167	0.145	0.191	0.171	0.149	0.196	0.298	0.270	0.327
175	0.1823	1.2	0.134	0.114	0.157	0.161	0.140	0.185	0.166	0.144	0.190	0.177	0.155	0.202	0.182	0.159	0.207	0.331	0.303	0.361
200	0.1823	1.2	0.145	0.125	0.168	0.189	0.166	0.214	0.175	0.153	0.200	0.191	0.168	0.217	0.215	0.191	0.242	0.360	0.331	0.390
50	0.2624	1.3	0.106	0.088	0.127	0.128	0.109	0.150	0.101	0.084	0.121	0.127	0.108	0.149	0.142	0.122	0.165	0.226	0.201	0.253
75	0.2624	1.3	0.120	0.101	0.142	0.152	0.131	0.176	0.140	0.120	0.163	0.153	0.132	0.177	0.173	0.151	0.198	0.277	0.250	0.306
100	0.2624	1.3	0.145	0.125	0.168	0.208	0.184	0.234	0.178	0.156	0.203	0.200	0.176	0.226	0.234	0.209	0.261	0.383	0.353	0.414
125	0.2624	1.3	0.185	0.162	0.210	0.214	0.190	0.240	0.204	0.180	0.230	0.237	0.212	0.264	0.261	0.235	0.289	0.474	0.443	0.505
150	0.2624	1.3	0.192	0.169	0.218	0.248	0.222	0.276	0.236	0.211	0.263	0.269	0.242	0.297	0.265	0.239	0.293	0.528	0.497	0.559
175	0.2624	1.3	0.229	0.204	0.256	0.275	0.248	0.303	0.257	0.231	0.285	0.299	0.271	0.328	0.325	0.297	0.355	0.595	0.564	0.625
200	0.2624	1.3	0.280	0.253	0.309	0.329	0.301	0.359	0.321	0.293	0.351	0.367	0.338	0.397	0.392	0.362	0.423	0.673	0.643	0.701
50	0.3365	1.4	0.119	0.100	0.141	0.154	0.133	0.178	0.141	0.121	0.164	0.153	0.132	0.177	0.161	0.140	0.185	0.303	0.275	0.332
75	0.3365	1.4	0.184	0.161	0.209	0.195	0.172	0.221	0.212	0.188	0.238	0.209	0.185	0.235	0.240	0.215	0.267	0.410	0.380	0.441
100	0.3365	1.4	0.221	0.196	0.248	0.253	0.227	0.281	0.260	0.234	0.288	0.288	0.261	0.317	0.302	0.274	0.331	0.580	0.549	0.610
125	0.3365	1.4	0.290	0.263	0.319	0.314	0.286	0.343	0.308	0.280	0.337	0.339	0.310	0.369	0.396	0.366	0.427	0.692	0.663	0.720

Table 1 Statistical power for all simulation settings. Point estimates, as well as Wilson confidence intervals are reported for all analysis approaches *(Continued)*

Size	Treatment	OR	T-test	CI lower	CI upper	T-test delta	CI lower	CI upper	I-test	CI lower	CI upper	I-test delta	CI lower	CI upper	ANCOVA	CI lower	CI upper	Transitional	CI lower	CI upper
150	0.3365	1.4	0.309	0.281	0.338	0.376	0.347	0.406	0.374	0.345	0.404	0.404	0.374	0.435	0.442	0.411	0.473	0.736	0.708	0.762
175	0.3365	1.4	0.329	0.301	0.359	0.399	0.369	0.430	0.396	0.366	0.427	0.434	0.404	0.465	0.463	0.432	0.494	0.800	0.774	0.824
200	0.3365	1.4	0.407	0.377	0.438	0.464	0.433	0.495	0.445	0.414	0.476	0.495	0.464	0.526	0.536	0.505	0.567	0.857	0.834	0.877
50	0.4055	1.5	0.162	0.140	0.186	0.178	0.156	0.203	0.196	0.173	0.222	0.190	0.167	0.215	0.210	0.186	0.236	0.392	0.362	0.423
75	0.4055	1.5	0.238	0.213	0.265	0.263	0.237	0.291	0.281	0.254	0.310	0.291	0.264	0.320	0.318	0.290	0.348	0.592	0.561	0.622
100	0.4055	1.5	0.302	0.274	0.331	0.354	0.325	0.384	0.366	0.337	0.396	0.390	0.360	0.421	0.392	0.362	0.423	0.737	0.709	0.763
125	0.4055	1.5	0.368	0.339	0.398	0.443	0.412	0.474	0.420	0.390	0.451	0.467	0.436	0.498	0.515	0.484	0.546	0.825	0.800	0.847
150	0.4055	1.5	0.397	0.367	0.428	0.509	0.478	0.540	0.467	0.436	0.498	0.546	0.515	0.577	0.583	0.552	0.613	0.891	0.870	0.909
175	0.4055	1.5	0.495	0.464	0.526	0.559	0.528	0.589	0.567	0.536	0.597	0.597	0.566	0.627	0.648	0.618	0.677	0.919	0.900	0.934
200	0.4055	1.5	0.530	0.499	0.561	0.616	0.585	0.646	0.598	0.567	0.628	0.669	0.639	0.697	0.706	0.677	0.733	0.967	0.954	0.976

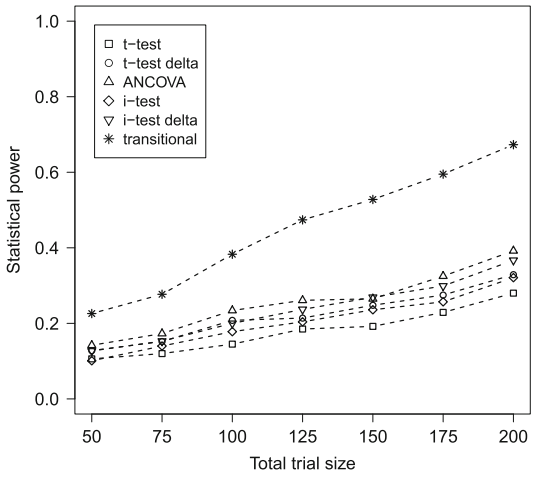

Fig. 1 Comparison of statistical power for the median treatment effect. The statistical power of all six approaches for treatment effect testing are plotted against total trial size (1:1 randomization) for the median simulated treatment effect $\beta_{\text{trt}} = 0.2624 = \log(1.3)$

controlling for baseline UEMS: $\widehat{\beta_{\text{trt}}} = -1.165$, $p-$value $= 0.307$.

i-test: No significant dependency between UEMS at 6 months and treatment arm: Z=0.553, $p-$value $= 0.58$.

i-test delta: No significant dependency between change in UEMS and treatment arm: Z=1.525, $p-$value $= 0.127$.

transitional: No significant shift in motor score probabilities associated with treatment arm: $\widehat{\beta_{\text{trt}}} = -0.197$, $p-$value $= 0.207$.

Summarizing, all six approached did not show significant results at the nominal level 0.05, but they all showed a tendency to less positive outcomes for patients in the treatment arm. This analysis is intended to give an example of the application for the proposed transitional ordinal model, but is not intended and should not be taken as a definitive conclusion about the value or outcome of the trial.

Discussion

The aim of this simulation study was to compare several approaches of testing for treatment effect in two-armed RCT in a neurological setting. We therefore simulated clinical trials with cervical SCI participants with specific levels of experimental conditions and tested for treatment effect with six different approaches. Routinely employed analysis approaches not only rely on strong assumptions about the properties of the endpoints being analysed, but were also outperformed in virtually all settings by the the proposed autoregressive transitional ordinal model for the analysis of UEMS.

Adding ordinal endpoints to form a single overall score is generally not valid

Common approaches to the analysis of UEMS and similar neurological endpoints are as the total sum of all motor scores $Y^*_{i,2} = \sum_{m=1}^{10} Y_{i,m,2}$ or as difference between two time points $Y^{**}_i = \sum_{m=1}^{10} Y_{i,m,2} - Y_{i,m,1}$.

Whether it is appropriate to combine a set of ordinal variables to generate a total score is usually not checked in neurology [37]. It should nonetheless be a requirement, as there are at least two strong assumptions related to the analysis of summed motor scores as a metric endpoint: unidimensionality and equal differences. Unidimensionality refers to the property of several scores to measure a single, common patient's characteristic. While there is some preliminary evidence that unidimensionality holds for UEMS [38], the opposite was reported for both the Functional Independence Measure FIM [39], the Spinal Cord Independence Measure SCIM [40], a situation which is very likely to be found in functional endpoints and Patients Reported Outcomes PRO. Equal differences imply that a unit change in motor scores represent exactly the same clinical change, independently of where the change took place on the scale (e.g. a change from 0 to 1 is assumed to be of the same magnitude as a change from 3 to 4 in motor scores), or of which key muscle are considered (the previous example is assumed to hold even when the changes took place on different key muscles, say e.g. one proximal and one distal from the lesion level).

The widely used method of adding up several ordinal endpoints to form a single overall score is therefore generally not valid with regard to the two assumptions exemplified above, and has been repeatedly reported in neurological and related physical functioning settings [39–44]. From a practical point of view, biased parameter estimates, as well as misleading associations and loss of power are some of the known consequences of assuming metric property for ordinal endpoints [21–23]. There is therefore a compelling need to embrace statistical models specifically designed for the analysis of complex ordinal endpoints.

RCT simulation

The proposed autoregressive transitional ordinal model is the first attempt in SCI to model and analyse a complex endpoint with a regression model which reflects its ordinal nature and takes into account important prognostic factors. The proposed model for the analysis of UEMS in cervical SCI patients outperformed all other approaches in virtually all settings. The sensibly lower statistical power achieved by commonly used approaches, in addition to their implicit assumptions, indicate that their use as default analysis methods in not justified.

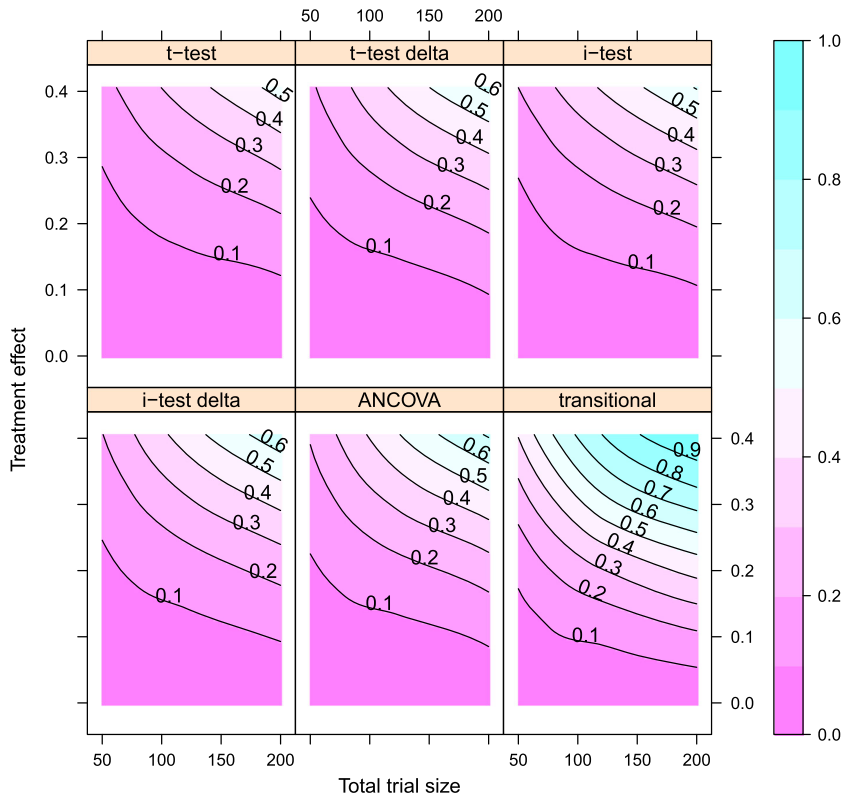

Fig. 2 Contour plots of statistical power for all simulation settings. The statistical power of all testing approaches is represented using loess smooth approximation. Contour curves visualize combinations of trial size and treatment effect with equivalent statistical power, which is reported as numerical value. The colour key differentiates regions of low statistical power (*violet*) from regions of high statistical power (*blue*)

Contrary to our expectations, a stratification of the t-test based on the Motor Level did not provide a discernible improvement in statistical power (Table 1). In fact, even though blocked independence tests showed a slightly higher power than their corresponding t-tests (Fig. 2), the gain in power was not such that their application as "ad hoc" solution resulted substantiated.

In terms of clinical interpretation of treatment effect estimates, we note that by applying the proposed model, the exponentiated treatment effect estimate $\widehat{\beta_{\text{trt}}}$ can be interpreted as the conditional odds ratio between the treatment and control trial arms, which is a common and accepted way of quantifying treatment effect in the clinical setting. Even when the proportional odds assumption is not fully met, it still provides an interpretable parameter that summarizes the treatment effect over all levels of the outcome [23]. In addition, the transitional model provides motor score probabilities for each combination of prognostic variables, making the direct comparison and visual representation of treated and untreated participants straightforward (see Fig. 3).

On the contrary, clear interpretation of the results produced by common approaches is precluded by summed scores of suppositional metric endpoints, providing little insight for trial scientists and clinicians. Importantly, small and possibly localized treatment effects, which are a hallmark of many neurological disorders, can be disentangled using ordinal approaches for motor scores, but become lost in the analysis of summed total scores.

Finally, our simulation showed (Table 1) that a statistical power of 80 %, which is a common goal for clinical trials planners, is reached by the ordinal model only for large trial size and large postulated treatment effects. As a total trial size of N=200 seems to currently represent the practical upper limit for conducting SCI trials, the statistical detection of an existing treatment effect seems to rely on a rather strong effect. Further improvements of the ordinal model will likely result in lowered requirements for treatment detection.

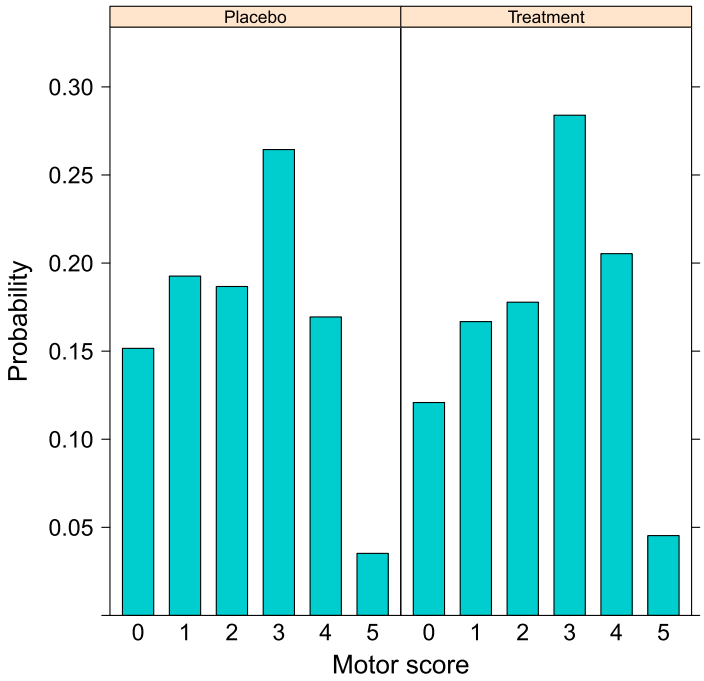

Fig. 3 Visualization of median treatment effect $\beta_{\text{trt}} = 0.2624 = \log(1.3)$. In contrast to all other analysis approaches, the transitional ordinal model allows to graphically represent shifts in motor score distributions for any constellation of relevant prognostic factors, permitting a much more detailed investigation of treatment effect. As illustrative example, represented is the distribution of motor score probabilities for participants in the control (*left* panel) and treatment arm (*right* panel). Lower scores became less, while higher score became more probable in the treatment arm. The treatment effect $\beta_{\text{trt}} = 0.2624 = \log(1.3)$ corresponds to an Odds Ratio of OR=1.3.The specific constellation of prognostics factor represented refers to a C8 key muscle, with a Motor Level C5 (x_{lev}=C5.-3), a baseline motor score of $y_{\text{base},i,m,1} = 1$, and an autoregressive component $y_{\text{auto},i,m-1,2} = 3$ for the motor score of the key muscle just above the one being reported

Revisiting a key SCI trial

To provide a concrete application of our approach, we analysed a subset of participants of the Sygen® trial [17, 35, 36]. Many ancillary analyses in the original publication were based on t-test and ANCOVA approaches and favoured the treatment group over placebo [17]. In particular, treated participants showed a faster initial recovery than control subjects, who nonetheless caught up at slightly later time points.

On the subsample of patients we considered, no one of the six approaches was significant at the conventional nominal level 0.05. Nonetheless, all approaches showed a tendency towards negative effect of treatment on the UEMS, meaning that treated patient showed on average a slightly worse recovery than patients in the control arm. Especially for the ordinal approach, the results imply that the odds of participants in the treatment group of achieving up to a given motor score were only $e^{\widehat{\beta_{\text{trt}}}} = 0.82$ times the odds of a participant with similar characteristics in the control arm, indicating a worse recovery for treated patients.

The negative estimate of treatment effect in cervical participants is rather unexpected. The observed unbalance toward more severe lesions in the treatment arm may explain at least in part these results, which nonetheless might be examined more closely to rule out potentially unintended detrimental effects. Nevertheless, we retain that generalizations of our results to the overall validity of the trial and its compound cannot be drawn.

Are summed overall scores not "good enough" ?

In our application, all six approaches presented delivered comparable results, namely statistically non-significant negative trends for participants in the treatment arm. One may therefore wonder what the added value of an ordinal approach like the proposed transitional ordinal model is. Briefly, routinely employed approaches based on summed overall scores imply:

- Unmet assumptions: adding ordinal endpoints to form a single overall score requires equal differences across all ordinal scales as well as unidimensionality.

Both assumptions are usually not further investigated [37], but the first can be rejected on medical reasons only, while the latter does not hold for several SCI endpoints (e.g. FIM [39], SCIM [40]).
- Flawed inference and estimation: known practical consequences of assuming metric property for ordinal endpoints are biased parameter estimates and misleading associations [21–23].
- Reduced statistical power: small and possibly localised effects are expected to be the hallmark of spinal cord injury rehabilitation strategies. The simulation reported provide evidence for a much lower capacity of approaches based on summed scores to detect existing treatment effects. Lower power also translates in higher requirement for trial participants.
- Unclear interpretation of treatment effect: a clear interpretation of treatment effect estimates as conditional OR, which can be visualised for each key muscle separately (see Fig. 3), is not possible for summed scores.
- Limited future extensions: future refinement of routinely employed approaches are strongly limited by the underlying, inappropriate analysis approach. Instead, ordinal approaches, which are based on a regression framework, easily accommodates for extensions (e.g. further prognostic factors, interactions, localised effects).

Concluding, from a theoretical point of view, routinely employed approaches have little scientific validity and have been replaced by more rigorous approaches. Even more importantly, they are also potentially misleading on practical terms. Our flexible model represents therefore an improved and pragmatic solution to the analysis of this type of complex ordinal endpoints.

Brain Injury: similar issues, similar solutions

We observe that most of the discussion points we raised link to the report by the International Mission on Prognosis and Clinical Trial Design in Traumatic Brain Injury TBI [45]. TBI is a related clinical field which faced very similar challenges, mainly related to the heterogeneity of the patient population, and had a similar history of clinical testing as SCI.

In fact, TBI also experienced a disappointing progression of clinical testing of treatment interventions in spite of extremely promising pre-clinical data and early phase trials. Maas et al. [45] reported that a key difficulty has been the inherent heterogeneity TBI subjects, and that the observed development was due, at least to some extent, to limitations in the trial designs and analyses. Both aspects have also been reported as hallmarks of SCI research.

Summarizing, The TBI Mission solicited the TBI community to [45]:

- provide details of the major baseline prognostic characteristics
- broaden inclusion criteria as much as is it compatible with the current understanding of the mechanisms of action of the intervention
- incorporate pre-specified covariate adjustment into the statistical analysis
- use an ordinal approach for the statistical analysis

A part from the first recommendation, which is mainly concerned with the way clinical studies are reported, the following three points regard the planning and especially the analysis of clinical trials in TBI, and are implemented in this publication. Selection of patients is based only on the initial Motor Level, which relates to the understanding of motor function. The proposed model (see Eq. 2) both include the most relevant covariates adjustment, namely baseline motor scores as well as motor lesion, and uses and ordinal approach for ordinal data based on the proportional odds model.

Latent variable models: an improved, readily available framework

More generally speaking, the statistical foundations of regression models for ordinal endpoints were developed more than 4 decades ago [46–48], and have ever since undergone a steady development. There is a huge body of literature pertaining to the analysis of ordinal variables, including Item Response Theory IRT and mixed-effects models for ordinal variables [49]. Despite this development, most clinical trials in neurology still rely on surpassed approaches [44], corroborating the negative trend of methodological errors related to the analysis of ordinal scales in medical research [50].

The proposed transitional ordinal model (Eq. 2) is an extension of the well known proportional odds model (e.g. [51]). The latter can be seen as an important special case within the IRT framework, and is closely related to the Rasch model [46]. All these statistical models are generally referred to as latent variable models, because they find application in situations where a set of ordinal variables are seen as indicators of a latent variable. This latent variable is the main interest of the analysis, and, although it cannot be measured directly, it can be inferred from the available ordinal variables. The latent variable approach seems both appropriate and appealing for applications in the clinical setting, and the transitional ordinal model proposed draws a concrete link from SCI to latent variable models. Further extensions of our approach can be tailored to the analysis of other endpoints such as functional assessments and PROs. In fact, the analysis of PRO, and the related trial powering based on Rasch models has recently received much attention [52, 53]. We believe that the transition from currently employed

analysis approaches to more sophisticated models within the readily available framework of latent variable models would represent a great scientific progression for the planning and analysis of complex neurological endpoints.

Conclusion

We propose an autoregressive transitional ordinal model for the analysis of a specific SCI endpoint which takes into account the complex ordinal nature of the endpoint under investigation and explicitly accounts for relevant prognostic factors. Superior statistical power in virtually all settings, combined with a clear clinical interpretation of treatment effect and widespread availability on commercial softwares, are strong arguments for clinicians and trial scientists to adopt, and further refine, the proposed approach.

Additional file

Additional file 1: International Standards for Neurological Classification of Spinal Cord Injury. (PDF 935 kb)

Abbreviations

EMSCI: European multicenter study about spinal cord injury; FIM: Functional independence measure; ICCP: International campaign for cures of spinal cord injury paralysis; IRT: Item response theory ISNCSCI: Int. standards for neurological classification of spinal cord injury; ML: Motor level; OR: Odds ratio; PRO: Patient reported outcomes; RCT: Randomized clinical trial; SCI: Spinal cord injury; SCIM: Spinal cord independece measure; TBI: Traumatic brain injury; UEMS: Upper Extremity Motor Scores

Acknowledgements

We appreciate the continuous assistance of René Koller with the EMSCI database.

Funding

LGT was partially financially supported by the International Foundation for Research in Paraplegia. The Foundation had no influence on any aspect of this publication.

Availability of data and materials

The datasets supporting the conclusions of this article are not publicly available. Interested researcher may apply for data access to the responsible organization, which is usually granted for research-only purposes. The R code implementing the simulation study is freely available (doi:http://dx.doi.org/10.5281/zenodo.47600).

Authors' contributions

LGT conceived the study, implemented the simulation and performed the analysis, and drafted the manuscript. JDS participated in the interpretation of the analyses and revision of the manuscript. AC participated in the interpretation of the analyses and revision of the manuscript. TH conceived the study, and participated in the simulation, interpretation, and drafting of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The data utilized in this study was extracted from the European Multicenter Study about Spinal Cord Injury (EMSCI, ClinicalTrials.gov Identifier: NCT01571531, www.emsci.org). All patients gave written informed consent. The ethical committee of the Canton of Zurich, Switzerland, has previously approved the EMSCI project, upon which this project is based, and the approval is also valid for any statistical analysis/re-analysis.

Author details

[1]Department of Biostatistics; Epidemiology, Biostatistics and Prevention Institute; University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland. [2]ICORD, University of British Columbia and Vancouver Coastal Health, Vancouver, Canada. [3]Spinal Cord Injury Center, Balgrist University Hospital, Zurich, Switzerland.

Received: 31 May 2016 Accepted: 19 October 2016
Published online: 08 November 2016

References

1. Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, Dodel R, Ekman M, Faravelli C, Fratiglioni L, Gannon B, Jones DH, Jennum P, Jordanova A, Jönsson L, Karampampa K, Knapp M, Kobelt G, Kurth T, Lieb R, Linde M, Ljungcrantz C, Maercker A, Melin B, Moscarelli M, Musayev A, Norwood F, Preisig M, Pugliatti M, Rehm J, Salvador-Carulla L, Schlehofer B, Simon R, Steinhausen HC, Stovner LJ, Vallat JM, den Bergh PV, van Os J, Vos P, Xu W, Wittchen HU, Jönsson B, Olesen J. Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol. 2011;21(10):718–79.
2. Andlin-Sobocki P, Jönsson B, Wittchen HU, Olesen J. Cost of Disorders of the Brain in Europe. Eur J Neurol. 2005;12:1–27.
3. Thuret S, Moon LDF, Gage FH. Therapeutic interventions after spinal cord injury. Nat Rev Neurosci. 2006;7(8):628–43.
4. Tator CH. Review of treatment trials in human spinal cord injury: issues, difficulties, and recommendations. Neurosurgery. 2006:957–987.
5. Hawryluk GW, Rowland J, Kwon BK, Fehlings MG. Protection and repair of the injured spinal cord: a review of completed, ongoing, and planned clinical trials for acute spinal cord injury: A review. Neurosurgical focus. 2008;25(5):14.
6. Liu K, Tedeschi A, Park KK, He Z. Neuronal Intrinsic Mechanisms of Axon Regeneration. Ann Rev Neurosci. 2011;34(1):131–52.
7. Schwab ME, Buchli AD. Drug research: plug the real brain drain. Nature. 2012;483(7389):267–8.
8. Ducker TB, Hamit HF. Experimental treatments of acute spinal cord injury. J Neurosurg. 1969;30(6):693–7.
9. Lammertse DP. Clinical trials in spinal cord injury: lessons learned on the path to translation. The 2011 International Spinal Cord Society Sir Ludwig Guttmann Lecture. Spinal Cord. 2012;51(1):2–9.
10. Fawcett JW, Curt A, Steeves JD, Coleman WP, Tuszynski MH, Lammertse D, Bartlett PF, Blight AR, Dietz V, Ditunno J, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. Spinal Cord. 2006;45(3):190–205.
11. Steeves JD, Lammertse D, Curt A, Fawcett JW, Tuszynski MH, Ditunno JF, Ellaway PH, Fehlings MG, Guest JD, Kleitman N, et al. Guidelines for the conduct of clinical trials for spinal cord injury (SCI) as developed by the ICCP panel: clinical trial outcome measures. Spinal Cord. 2006;45(3):206–21.
12. Tuszynski MH, Steeves JD, Fawcett JW, Lammertse D, Kalichman M, Rask C, Curt A, Ditunno JF, Fehlings MG, Guest JD, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP Panel: clinical trial inclusion/exclusion criteria and ethics. Spinal Cord. 2006;45(3):222–31.
13. Lammertse D, Tuszynski MH, Steeves JD, Curt A, Fawcett JW, Rask C, Ditunno JF, Fehlings MG, Guest JD, Ellaway PH, et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: clinical trial design. Spinal Cord. 2006;45(3):232–42.
14. Sorani MD, Beattie MS, Bresnahan JC. A Quantitative Analysis of Clinical Trial Designs in Spinal Cord Injury Based on ICCP Guidelines. J Neurotrauma. 2012;29(9):1736–46.
15. Bracken MB, Shepard MJ, Collins WF, Holford TR, Young W, Baskin DS, Eisenberg HM, Flamm E, Leo-Summers L, Maroon J, Marshall LF, Perot PL, Piepmeier J, Sonntag VKH, Wagner FC, Wilberger JE, Winn HR. A randomized, Controlled Trial of Methylprednisolone or Naloxone in the

Treatment of Acute Spinal-Cord Injury - Results of the Second National Acute Spinal Cord Injury Study. New England J Med. 1990;322(20): 1405–11.
16. Hansebout RR, Blight AR, Fawcett S, Reddy K. 4-Aminopyridine in chronic spinal cord injury: a controlled, double-blind, crossover study in eight patients. J Neurotrauma. 1993;10(1):1–18.
17. Geisler FH, Coleman WP, Grieco G, Poonian D, Group SS. The Sygen®multicenter acute spinal cord injury study. Spine. 2001;26(24S): 87–98.
18. Lammertse DP, Jones LAT, Charlifue SB, Kirshblum SC, Apple DF, Ragnarsson KT, Falci SP, Heary RF, Choudhri TF, Jenkins AL, Betz RR, Poonian D, Cuthbert JP, Jha A, Snyder DA, Knoller N. Autologous incubated macrophage therapy in acute, complete spinal cord injury: results of the phase 2 randomized controlled multicenter trial. Spinal Cord. 2012;50(9):661–71.
19. Casha S, Zygun D, McGowan MD, Bains I, Yong VW, John Hurlbert R. Results of a phase II placebo-controlled randomized trial of minocycline in acute spinal cord injury. Brain. 2012;135(4):1224–36.
20. Agresti A. Analysis of Ordinal Categorical Data, 2nd ed. Series in Probability and Statistics. Hoboken: Wiley; 2010.
21. Winship C, Mare RD. Regression models with ordinal variables. Am Sociol Rev. 1984:512–25.
22. Hastie TJ, Botha JL, Schnitzler CM. Regression with an ordered categorical response. Stat Med. 1989;8(7):785–94.
23. Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. J Clin Epidemiol. 1997;50(1):45–55.
24. Kirshblum SC, Waring W, Biering-Sorensen F, Burns SP, Johansen M, Schmidt-Read M, Donovan W, Graves DE, Jha A, Jones L, Mulcahey MJ, Krassioukov A. Reference for the 2011 revision of the international standards for neurological classification of spinal cord injury. J Spinal Cord Med. 2011;34(6):547–54.
25. Rudhe C, van Hedel HJA. Upper Extremity Function in Persons with Tetraplegia: Relationships Between Strength, Capacity, and the Spinal Cord Independence Measure. Neurorehabil Neural Repair. 2009;23(5): 413–21.
26. Coleman W. Injury severity as primary predictor of outcome in acute spinal cord injury: retrospective results from a large multicenter clinical trial*1. Spine J. 2004;4(4):373–8.
27. Tanadini LG, Hothorn T, Jones LA, Lammertse DP, Abel R, Maier D, Rupp R, Weidner N, Curt A, Steeves JD. Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury. Neurorehabil Neural Repair. 2015;29(9):867–77.
28. Hothorn T, Hornik K, van de Wiel MA, Winell H, Zeileis A. Coin: Conditional Inference Procedures in a Permutation Test Framework. 2015. http://CRAN.R-project.org/package=coin.
29. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. A Lego System for Conditional Inference. Am Stat. 2006;60(3):257–63.
30. Ripley B. MASS: Support Functions and Datasets for Venables and Ripley's MASS. 2015. http://CRAN.R-project.org/package=MASS.
31. Nenables WN, Ripley BD. Modern Applied Statistics With S, 4th ed. New York: Springer; 2002. http://www.stats.ox.ac.uk/pub/MASS4.
32. Kennedy P, Cade B. Randomization tests for multiple regression. Commun Stat - Simulation Comput. 1996;25(4):923–36.
33. Parhat P, Rosenberger WF, Diao G. Conditional Monte Carlo randomization tests for regression models. Stat Med. 2014;33(18): 3078–088.
34. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2015. https://www.R-project.org/.
35. Geisler FH, Coleman WP, Grieco G, Poonian D, Group SS, et al. Recruitment and early treatment in a multicenter study of acute spinal cord injury. Spine. 2001;26(24S):58–67.
36. Geisler FH, Coleman WP, Grieco G, Poonian D, Group SS. Measurements and recovery patterns in a multicenter study of acute spinal cord injury. Spine. 2001;26(24S):68–86.
37. Hobart J. Rating scales for neurologists. J Neurol Neurosurg Psychiatry. 2003;74(suppl 4):22–6.
38. Furlan JC, Fehlings MG, Tator CH, Davis AM. Motor and Sensory Assessment of Patients in Clinical Trials for Pharmacological Therapy of Acute Spinal Cord Injury: Psychometric Properties of the ASIA Standards. J Neurotrauma. 2008;25(11):1273–1301.
39. Ravaud JF, Delcey M, Yelnik A, et al. Construct validity of the functional independence measure (FIM): Questioning the unidimensionality of the scale and the "value" of FIM scores. Scand J Rehabil Med. 1999;31(1):31–42.
40. Catz A, Itzkovich M, Tesio L, Biering-Sorensen F, Weeks C, Laramee MT, Craven BC, Tonack M, Hitzig SL, Glaser E, et al. A multicenter international study on the Spinal Cord Independence Measure, version III: Rasch psychometric validation. Spinal Cord. 2007;45(4):275–91.
41. McHorney CA, Haley SM, Ware JEJ. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II, Comparison of relative Precision Using Likert and Rasch Scoring Methods. J Clin Epidemiol. 1997;50(4):451–61.
42. Fink P, Ewald H, Jensen J, Sorensen L, Engberg M, Holm M, Munk-Jorgensen P. Screening for somatization and hypochondriasis in primary care and neurological in-patients: a seven-item scale for hypochondriasis and somatization. J Psychosomatic Res. 1999;46(3): 261–73.
43. Luther SL, Kromrey J, Powell-Cope G, Rosenberg D, Nelson A, Ahmed S, Quigley P. A Pilot Study to Modify the SF-36v Physical Functioning Scale for Use With Veterans With Spinal Cord Injury. Arch Phys Med Rehabil. 2006;87(8):1059–66.
44. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurology. 2007;6(12):1094–1105.
45. Maas AI, Steyerberg EW, Marmarou A, McHugh GS, Lingsma HF, Butcher I, Lu J, Weir J, Roozenbeek B, Murray GD. IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. Neurotherapeutics. 2010;7(1):127–34.
46. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests vol. Copenhagen: Paedagogiske Institut; 1960.
47. McKelvey RD, Zavoina W. A statistical model for the analysis of ordinal level dependent variables. J Math Sociol. 1975;4:103–20.
48. McCullagh P. Regression models for Ordinal Data. J Royal Stat Soc. 1980;42(2):109–42.
49. Mehta PD, Neale MC, Flay BR. Squeezing Interval Change From Ordinal Panel Data: Latent Growth Curves With Ordinal Outcomes. Psychol Methods. 2004;9(3):301–33.
50. Forrest M, Andersen B. Ordinal scale and statistics in medical research. Br Med J (Clinical research ed.) 1986;292(6519):537.
51. Ananth CV, Kleinbaum DG. Regression Models for Ordinal Responses: A Review of Methods and Applications. Int J Epidemiol. 1997;26(6):1323–33.
52. McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, Weir J, Maas AIR, Murray GD. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. Clinical Trials. 2010;7(1):44–57.
53. Hardouin JB, Blanchin M, Feddag ML, Néel TL, Perrot B, Sébille V. Power and sample size determination for group comparison of patient-reported outcomes using polytomous Rasch models. Stat Med. 2015;34(16): 2444–455.

PAPER IV

Addressing limitations of rating scales and their analysis in spinal cord injury under the unifying framework of latent variable modeling

Lorenzo G. Tanadini, Armin Curt, Irini Moustaki

Submitted manuscript.

Addressing limitations of rating scales and their analysis in spinal cord injury under the unifying framework of latent variable modeling

Lorenzo G. Tanadini[a,b,*], Armin Curt[c], Irini Moustaki[b]

[a] *Dept. of Biostatistics, University of Zurich*
[b] *Dept. of Statistics, London School of Economics*
[c] *Spinal Cord Injury Center, Balgrist University Hospital*

Abstract

Background. Despite the ubiquity of multiple-item rating scales used to measure health outcomes in patients with neurological disorders, current analysis approaches present several limitations and represent a major weakness of the translational process.

Methods. We conducted a retrospective analysis of prospectively collected longitudinal data of patients who suffered a spinal cord injury with the goal to assess metric properties and dimensionality of a multiple-item neurological rating scale. In addition, we propose an approach for the longitudinal modeling of neurological recovery after injury. All analyses were performed with the unifying framework of latent variable modeling and involved latent trait analysis and latent curve modeling techniques.

Findings. The analysis provided evidence that the multiple-item primary endpoint Upper Extremity Motor Score behaves as an ordinal scale, but, due to varying loadings across key muscles, its analysis as a total sum score will be insufficient as a measure of the attribute. In addition, unidimensionality holds only for a subset of the initially selected patient population. We found strong evidence that the key muscles composing the Upper Extremity Motor Score on the left and right body sides measure the same latent construct, allowing important simplification in following analysis steps. The longitudinal modeling reported a strong negative correlation between intercept and slope terms.

Interpretation. Our comprehensive study represents a concrete solution for the analysis of a complex neurological endpoint that avoids commonly used, but flawed approaches. It is further intended to act as a template for the assessment of metric properties and analysis of multiple-item rating scales across medical disciplines.

*Corresponding author
Email address: lorenzo.tanadini@uzh.ch (Lorenzo G. Tanadini)

Preprint submitted to Elsevier *May 16, 2017*

Funding. Swiss National Science Foundation (Grant P1ZHP3_158783), Janggen-Phoen Foundation.

Keywords:

- complex ordinal endpoints
- metric properties and dimensionality
- longitudinal analysis
- latent growth curve models
- testing treatment efficacy

Research in context.

Evidence before this study Two review articles comprehensively summarize the scientific evidence considered before this study. Hobart (2007) reported that methodological limitations of rating scales are an important cause of clinical failure of very promising therapeutic approaches in neurology. An historical revision of key studies in the field of spinal cord injury by Lammertse (2012) reveals that multiple-item rating scales are consistently analyzed as single overall summed scores. Nonetheless, those approaches have been shown to be inappropriate in many aspects.

Added value of this study Our comprehensive investigation of a frequently used neurological endpoint in spinal cord injury trials assessed metric properties and proposed a longitudinal endpoint analysis based on the unifying framework of latent variable modeling. Our study exemplifies theoretical and practical drawbacks of commonly used approaches, and proposes a concrete solution based on a much more rigorous and sophisticated statistical framework.

Implications of all the available evidence Our findings support the concern that methodological limitations of rating scales in neurology are a major factor contributing to the failure of many clinical studies in this field. Our results provide a concrete solution to investigating metric properties and analyzing a primary endpoint generated by multiple-item rating scales. While basing our analysis on a function-based neurological endpoint often applied in spinal cord injury, the content of our paper is highly relevant to all neurological disciplines. Not only are multiple-item rating scales omnipresent in all fields of neurology, but our analyses are also intended to act as a template for the assessment of this type of endpoints across neurology.

1. Introduction

Neurological research is responsible for the investigation of many devastating disorders, with the ultimate goal of delivering effective treatments to patients. Many discoveries from basic research and clinical investigation have entered the translational process, which seems to be based on optimal premises due to the acceleration in understanding of brain plasticity and the economic potential of any drug that obtained approval from regulatory agencies. In fact, with yearly health costs for European society estimated at almost 400 billion € [1], brain-related disorders are a greater socio-economic burden than cancer, cardiovascular diseases and diabetes combined [2]. Despite the obvious economic interest, health insurers and drug companies have nonetheless withdrawn from neuroscience as a consequence of the large number of failed trials [3].

The question therefore arises as why very promising approaches have failed to deliver treatments in humans. In line with others [4], we maintain that major weaknesses of the translational process are to be located in the methods used to test for efficacy of treatment in rating scales used to measure health outcomes in patients. There are, in fact, two often overlooked limitations of multiple-item rating scales that are particularly relevant. Firstly, the ordinal numbers generated by rating scale items are characterized by arbitrary integers solely establishing a ranking. This implies that the difference between two following ranks is by no means bound to be equivalent in clinical terms, and a given difference in total score usually does not have the same meaning across the scale range [5]. Secondly, instead of a single, unidimensional health domain, many neurological multiple-item rating scales measure a combination of several domains. While this may seem to be a rather theoretical issue far away from clinical practice, it means having a scale that measures "length at one end, weight in the middle, and volume at the other end" [5].

Nonetheless, multiple-item rating scale measurements are usually taken at face values and handled as continuous, interval-scaled measurements and often combined in a total summed score before analysis. Such an approach has been shown to be inappropriate in a number of aspects [6], with consequences such as biased parameter estimates, misleading associations and loss of power [7, 8, 9].

Despite all these issues, ordinal measurements generated by multiple-item rating scales are often the only type of data which clinical assessments provide and thus what neurologists have to work with [5]. To overcome shortcomings of currently employed approaches, we assess metric properties and dimensionality of a multiple-item neurological rating scale and we propose an approach for the longitudinal modeling of neurological recovery based on latent variable modeling. This not only places it in a much more rigorous and sophisticated inferential framework, it also provides a template applicable to similar endpoints across medical fields.

2. Methods

The objective of this study was to propose an approach to the analysis of complex ordinal endpoints which overcomes drawbacks of common approaches in neurology. We report a comprehensive analysis of a multiple-item rating scale particularly relevant in spinal cord injury research, providing a widely applicable template for the assessment of its metric properties and for its longitudinal analysis under the unifying framework of latent variables.

Patient population and study design. For our retrospective analysis of prospectively collected longitudinal data, we utilised records of N=782 de-identified individuals extracted from the European Multicenter Study about Spinal Cord Injury (EMSCI, ClinicalTrials.gov). EMSCI tracks, in a highly standardized manner, the functional and neurological recovery of patients at 1, 4, 12, 24, 48 weeks after spinal cord injury. All examinations were performed by trained staff according to the International Standards for Neurological Classification of Spinal Cord Injuries [10]. The Ethical Committee of the Canton of Zurich, Switzerland, approved the EMSCI project and any statistical analysis thereof.

Primary endpoint. The primary focus of our analyses was the Upper Extremity Motor Scores (UEMS), which is a particularly relevant neurological endpoint in the field of spinal cord injury [11, 12, 13]. UEMS measures the muscle contraction force for 10 key muscles on the arms and hands. On each body side, the elbow flexors (C5), wrist extensors (C6), elbow extensors (C7), finger flexors (C8), and little finger abductors (T1) are tested. Each key muscle is rated on a 6-point rating scale with the following muscle function grading: "total paralysis" (score 0), "palpable or visible contraction" (score 1), "active movement, full range of motion (ROM) with gravity eliminated" (score 2), "active movement, full ROM against gravity" (score 3), "active movement, full ROM against gravity and moderate resistance in a muscle specific position" (score 4), "normal active movement, full ROM against gravity and full resistance in a functional muscle position expected from an otherwise unimpaired person" (score 5).

Statistical analysis. Latent variable modeling, of which factor analysis is the oldest and most widely used, finds applications in all situations where the variable of major interest cannot be measured directly [14]. Instead, a set of observable variables called indicators (the key muscles) are measured and used to make inference about the latent variable of interest (neurological status). For the specific neurological settings outlined here, where all indicators are measured on rating scales, and the latent variables inferred are supposedly metrical, the techniques are collectively referred to as Latent Trait Analysis (LTA). LTA methods have a common mathematical formulation described in Equations 1 - 3. Let suppose that the observed item scores (e.g. the rating of the ten key muscles composing the UEMS) are denoted by x_1,...,x_p, and that there are m_i ordered categories for item i. Let $\pi_{i(s)}(\mathbf{f})$ be the probability that, given the latent variables $\mathbf{f}$, a measurement delivers response category s to item i. The category response function $\pi_{i(s)}(\mathbf{f})$ (Equation 1) can be expressed in terms of

difference between adjacent cumulative response functions (Equation 2), and thus the model is written as indicated in Equation 3.

$$\pi_{i(s)}(\mathbf{f}) = \gamma_{i(s)}(\mathbf{f}) - \gamma_{i(s-1)}(\mathbf{f}) \tag{1}$$

$$\gamma_{i(s)}(\mathbf{f}) = P(x_i \leq s \mid \mathbf{f}) = \pi_{i(1)}(\mathbf{f}) + \pi_{i(2)}(\mathbf{f}) + ... + \pi_{i(s)}(\mathbf{f}) \tag{2}$$

$$\log\left[\frac{\gamma_{i(s)}(\mathbf{f})}{1-\gamma_{i(s)}(\mathbf{f})}\right] = \alpha_{i(s)} - \sum_{j=1}^{q} \alpha_{ij} f_j \tag{3}$$

where $s = 1, ..., m_i$, $i = 1, ..., p$ and $f_j \sim \mathrm{N}(0,1)$ for $j = 1, ..., d$ are independent standard normal distributions. The model represented in Equation 3 is also known as the proportional odds model [15]. The factor loadings α_{ij} quantify how the probability distributions of key muscle scores are influenced by the position of a given patient on the latent variables f_j (e.g. his neurological status).

Metric properties and dimensionality Firstly, we performed an exploratory nominal LTA [16] to assess whether the key muscles behave as intended as a rating scale. This analysis was conducted using all SCI patients assessed at one week after injury ($t = 1$), and for left and right body parts separately. As the factor loadings show the increase in the odds of falling into a response category as the position of a given patient on the latent variable increases, a situation in which the ordered factor loadings increased in magnitude would be an indication that the items are indeed measured on an ordinal scale [14]. In addition, the standardized loadings provide insight into the validity of summing all motor scores in an unweighted total score. In fact, when the standardized loadings vary between items, the analysis based on the unweighted sum of all motor scores (e.g. total UEMS) is invalid [14].

Secondly, we performed an ordinal LTA to assess the dimensionality of UEMS. A desirable property of rating scales is that all items are indicators of a single, unidimensional latent health domain. The dimensionality is assessed by fitting LTA models with an increasing number of latent dimensions, which are then compared in terms of model fit to determine how many latent variables are necessary to best reproduce the observed data. Model fit was assessed by looking at the fits to the one- and two-way margins, where standardized Chi-square type residuals greater than 4 would indicate poor fit [14]. In addition, model selection criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) provided additional elements of guidance for model selection [17]. This analysis was conducted for all SCI patients, separately for each time point and for each body side. Given that UEMS measures five key muscles on each body side, we fitted 1- and 2-dimensional LTA models.

Thirdly, we ran a Confirmatory Factory Analysis (CFA) to assess whether key muscles on the right and left body sides are measuring the same latent construct. A two-factor CFA was fitted to the left and right body sides, with

indicators from each side loading on a separate latent variable, which were allowed to correlate. A high correlation of the left and right latent variables would be a strong indication that both sets of 5 key muscles are indeed measuring the same latent construct. This analysis was conducted for all cervical SCI participants, separately for each time point.

Longitudinal modeling The longitudinal evolution of the neurological status measured by UEMS, and indeed all multiple-item rating scales, can also be modeled within the framework of latent variables by applying a Latent Growth Curve Model (LGCM) [18]. As for previous analysis, the key muscles are used as indicators of the underlying continuous neurological status. In addition, the inferred multiple measurements of neurological status at different time points are seen themselves as indicators of latent variables describing the intercept and slope components of the trajectory modelled as graphically represented in Figure 1. By applying LGCM, we achieve a flexible model with patient-specific trajectories of neurological recovery over time [19].

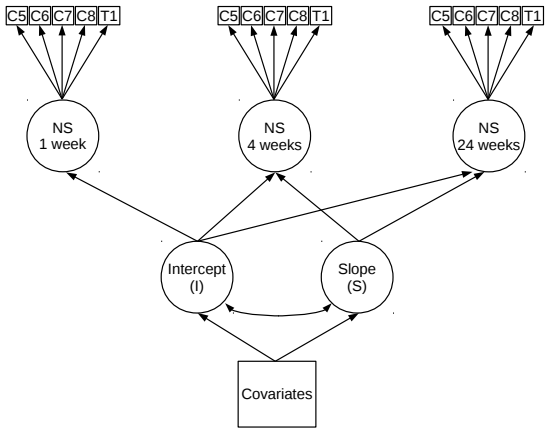

Figure 1: Simplified path diagram representing the dependencies among key muscles as indicators (top boxes), the latent variables quantifying the neurological status (NS) at different time points (top circles), which act themselves as indicators for further latent variables coding for the intercept (I) and slope (S) components of the longitudinal model (bottom circles). Baseline covariates were allowed to directly influence both intercept and slope (bottom box).

The longitudinal modeling of UEMS was conducted including all cervical SCI patients at 1, 4 and 24 weeks after injury. Only at these time points, the extent of missing data on all key muscles was marginal (12%, 11%, 7% respectively) and the condition of data missing at random still plausible. The missing key muscles are treated with the direct Maximum Likelihood (ML) approach, which

maintains the ML properties under the assumption of missing at random [20]. The availability of three time points allowed only the formulation of models for a linear recovery. Nonetheless, the focus of this manuscript is the application of latent variables models to neurological situations requiring it, rather than the detailed modeling of the recovery shape.

Routinely available covariates collected at baseline such as gender, height, age at injury, year of injury, and motor level (right body side) were allowed to directly influence both the intercept and slope components of the LGCM in a classical regression fashion (see lower part Figure 1). In future studies, further variables such as treatment arm can be added in a straightforward manner, providing a test for treatment effect in the longitudinal analysis of neurological endpoints. The utility of LGCM in treatment comparisons and their ability to address possible differences in the initial conditions of trial participants in different trial arms have been reported before [19].

All computations were performed with M*plus* (version 7.11). Nominal LTA was performed with LAMI (freely available at LAMI). Input files are provided as supplementary web material.

Role of the funding sources. The entities providing financial support had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

3. Results

Metric properties and dimensionality. Firstly, we performed an exploratory nominal Latent Trait Analysis (LTA) to assess whether the Upper Extremity Motor Score (UEMS) behaves as intended as a rating scale. Table 1 shows the standardized loadings for the left and the right body sides, which were analyzed independently at 1 week after injury.

For both sides of the body, the estimated loadings of each key muscle show a pattern of increasing loadings with increasing category, implying an increase in the odds of falling into a given category compared to the reference category with an increase in the patient position on the latent variable. In addition, the estimated loadings $st\alpha_{ij(s)}$ vary between 0.5100 to 0.9962 and are therefore not of similar magnitude. This finding does not support the use of the sum score as the unit of analysis.

Secondly, we performed an ordinal LTA to assess the dimensionality of UEMS. A desirable property of rating scales is that all indicators load on a single, unidimensional latent health domain. The goodness-of-fit of LTA models with increasing number of latent traits was assessed by looking at the fits to the one- and two-way margins, where standardized residuals greater than 4 would indicate poor fit [14].

An important number of residuals indicated poor model fit. The majority of standardized residuals that needed closer attention regarded category 5 of the motor scores. Especially the combinations of scores 5 and 5, independently of

Table 1: Nominal Latent Trait Analysis: standardized factor loadings of the one-factor model for all patients at 1 week after injury.

Key muscle	Score category (reference is 0)	Left body side	Right body side
C5	1	0.5100	0.8834
C5	2	0.8609	0.9057
C5	3	0.8845	0.9368
C5	4	0.9510	0.9580
C5	5	0.9865	0.9863
C6	1	0.7190	0.5553
C6	2	0.8452	0.8589
C6	3	0.9120	0.9397
C6	4	0.9761	0.9771
C6	5	0.9951	0.9950
C7	1	0.5795	0.6406
C7	2	0.8768	0.8979
C7	3	0.9481	0.9493
C7	4	0.9818	0.9829
C7	5	0.9956	0.9957
C8	1	0.8771	0.8281
C8	2	0.9371	0.8715
C8	3	0.9750	0.9710
C8	4	0.9835	0.9852
C8	5	0.9959	0.9962
T1	1	0.8990	0.8579
T1	2	0.9571	0.9423
T1	3	0.9685	0.9580
T1	4	0.9843	0.9861
T1	5	0.9954	0.9954

key muscle, produced the largest positive residuals in the one-factor model, indicating that the misfit is predominantly due to a unexpectedly large frequency of motor scores 5 compared to the model-based expectations. Even though the number of standardized residuals indicating poor fit was less severe for data at 4 and 24 weeks, the extent of misfit for EFA at 1 week after injury, and especially the one-factor model, prompted us to restrict the analysis to patients who suffered a spinal cord injury in the cervical region. Cervical patients represent a clinically relevant cohort, and we assumed that their high level of lesion would mitigate the excess of maximal muscle score leading to the severe misfit reported in Table 2. We therefore ran an ordinal LTA including only cervical patients to assess the dimensionality of UEMS, which is reported in Table 3.

Also for this model, the estimated loadings varied across items and therefore the total sum score will be insufficient as a measure of the attribute.

Thirdly, we ran a Confirmatory Factory Analysis (CFA) to assess whether there is indication that key muscles on the right and left body sides are mea-

Table 2: EFA for UEMS (left body side), all patients at Week 1 after injury (N=689): count of standardized residuals (z-scores) > |4| from the two-way margins. Counts concerning UEMS score 5 are reported in brackets.

	One-factor model					Two-factor model			
Items	C6	C7	C8	T1	Items	C6	C7	C8	T1
C5	3 (3)	2 (2)	3 (3)	3 (3)	C5	0	0	0	0
C6		4 (4)	2 (2)	2 (2)	C6		0	1 (0)	0
C7			1 (1)	1 (1)	C7			1 (1)	1 (0)
C8				2 (1)	C8				0

Table 3: One-factor EFA for UEMS (left body side), cervical patients at 1, 4, 24 weeks after injury: count of standardized residuals (z-scores) > |4| from the two-way margins of the one-factor model. Counts concerning UEMS score 5 are reported in brackets.

	1 week (N=195)				4 weeks (N=223)				24 weeks (N=218)			
Items	C6	C7	C8	T1	C6	C7	C8	T1	C6	C7	C8	T1
C5	0	0	0	0	0	0	0	0	1 (1)	0	0	0
C6		0	0	0		0	0	0		0	1 (0)	0
C7			0	0			1 (1)	0			1 (1)	0
C8				0				0				0

suring the same latent construct. A two-factor CFA model was fitted to the left and right body side, with indicators from each side loading on a separate latent variable, which were allowed to correlate. A high correlation of the left and right latent variables would be a strong indicator that both sets of 5 key muscles are indeed measuring the same latent construct. Correlation analyses were conducted for all cervical participants, separately for each time point, and delivered correlation coefficients (S.E.) of 0.936 (0.019) at 1 week, 0.915 (0.023) at 4 weeks, and 0.931 (0.022) at 24 weeks.

Longitudinal modeling. The longitudinal development of UEMS was modelled within the framework of latent variables by applying a Latent Growth Curve Model (LGCM) as depicted in Figure 1. The multiple measurements at different time points are considered indicators for further latent variables describing the intercept and slope components of the trajectory. The application of LGCM with categorical indicators requires measurement invariance across time points. The patterns of factor loadings are reported in Table 4 and are in line with the assumptions of LGCM.

As the number of thresholds for key muscle C5 on both sides of the body varied across time points, we decided to drop this indicator from further analyses. Within the framework of latent variable modeling, one may opt to include the key muscle C5 in the analysis by constraining it to be invariant, but this goes beyond the scope of this manuscript. Parameter estimates for latent growth factors as well as significant regression coefficients are presented in Table 5. There was a strong negative correlation between intercept and slope terms, meaning that participants with an initial high neurological status did not recover much

Table 4: Factor loadings of the 2-factor CFA model for cervical patients at different time points, one latent factor for each body side

	1 week		4 weeks		24 weeks	
	Right	Left	Right	Left	Right	Left
C5	0.757	0.856	0.520	0.510	0.731	0.481
C6	1.849	1.967	1.582	1.710	1.404	1.346
C7	3.619	3.338	3.282	3.491	2.640	2.805
C8	10.328	13.481	8.091	7.494	8.327	7.835
T1	6.994	7.964	7.374	6.092	6.072	7.129

over time (low slope), and vice-versa. Also, the initial Motor Level seemed to have a rather strong effect on the intercept term, which describes the inital neurological status.

Table 5: Latent curve model estimates with covariates influencing the latent growth factors intercept (I) and slope (S).

Parameter	Estimate	Standard Error	Two-tailed P-values
Intercept I	0	0 (fixed)	
Variance I	16.445	3.967	0.000
Intercept S	205.575	159.019	0.196
Variance S	0.879	0.299	0.003
Correlation S, I	-0.788	0.310	0.011
Regression coefficient	**Estimate**	**Standard error**	**Two-tailed P-values**
Age at injury on I	0.048	0.021	0.023
Motor level on I	3.171	0.421	0.000
Motor level on S	-0.450	0.141	0.001

4. Discussion

Complex ordinal endpoints generated by multiple-item rating scales are ubiquitous in neurology, and often represent the only data format which clinical assessments provide [5]. While the issue related to their inappropriate statistical analysis is not new in medicine [21], it has grown to such an extent to be considered a major weakness of the translational process [4]. In Spinal Cord Injury (SCI), virtually all clinical assessments produce complex ordinal data [10], but are usually analyzed in ways that discard their ordinal nature [11, 22, 23, 24, 25].

The results of our nominal Latent Trait Analysis (LTA) (see Table 1) provided support for the ordinality of the key muscles, a property which has been so far assumed without further investigation. Further, the observed pattern of larger loading within key muscles is in line with expectations, as higher scores are assigned to patients able to perform more complex motor tasks [10]. No general pattern could be recognized among key muscles. The lack of a clear pattern is probably due to the initial inclusion of patients with different lesion

levels, as the relevance of the motor level for motor recovery has been reported before [26, 27]. The same analysis revealed that standardized loadings varied between key muscles, making commonly employed analysis approaches based on the (unweighted) total sum of all motor scores an insufficient or even invalid measure of the neurological attribute.

Ordinal LTA prompted us to restrict our analysis to participants with a cervical motor lesion in order to assure unidimensionality of the Upper Extremity Motor Scores (UEMS) (see Table 2-3). While some SCI studies focused on the clinically relevant subset of cervical patients, many did not (e.g. 11, 24). All these results provide strong evidence against the default application of total summed UEMS indiscriminately to all patients included into a clinical study.

The Latent Growth Curve Model (LGCM) (see Figure 1) was fitted in order to demonstrate the need for better modeling strategies for complex ordinal endpoints. The availability of three time points allowed only the formulation of models for a linear recovery, which we acknowledge is a simplification [28]. Nonetheless, the focus of this manuscript is the application of latent variables models to neurological situations requiring it, rather than the detailed modeling of the recovery shape. In line with clinical expectations [26, 27], the model suggested that the initial motor level is a relevant prognostic factor that should be accounted for in the analysis of UEMS. The other significant prognostic factor Age at injury played only a marginal role. Instead, the negative correlation between the intercept and slope terms seems to suggest a rather strong trade-off between initial level of injury and recovery, which may be indicative of further issues concerning the measurement tool UEMS as whole.

5. Conclusion

The failure of any clinical study can always be attributed to multiple causes. It is, nonetheless, remarkable how often the choice of complex ordinal endpoints and their statistical analyzes are reported among those reasons [29, 30]. Despite the fact that the statistical foundations of regression models for ordinal endpoints were developed several decades ago [31, 32, 33] and have undergone a steady development, a large number of clinical trials in neurology still rely on surpassed approaches [4]. The statistical analyses performed here aimed at assessing the metric properties and longitudinal evolution of a common primary endpoint in Spinal Cord Injury. The unifying framework of latent variable modeling adopted is intended to replace commonly used, but flawed analysis approaches. The progression reported can be used as a template for the analysis of any multiple-item rating scale across medical fields. Although the mathematical foundation of those methods can be quite challenging [4], they provide a huge potential to change the face of health outcomes measurement [5] by placing the analysis of complex ordinal endpoints within a much more rigorous and sophisticated inferential framework.

6. Contributions

LGT conceived the study, performed the analysis, and drafted the manuscript. AC provided access to the data and participated in the revision of the manuscript. IM supervised analyses and interpretation, and revised the manuscript. All authors read and approved the final version of the manuscript.

7. Conflict of interest

The authors declare that there are no conflicts of interest.

8. Acknowledgments

LGT acknowledges financial support by the Swiss National Science Foundation (Grant P1ZHP3_158783) and by the Janggen-Phoen Foundation. We appreciate the continuous assistance of René Koller with the EMSCI database, and the EMSCI network for granting access to their data.

References

[1] P. Andlin-Sobocki, B. Jönsson, H.-U. Wittchen, Jes Olesen, Cost of Disorders of the Brain in Europe, European Journal of Neurology 12 (Suppl. 1) (2005) 1–27.

[2] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni, B. Gannon, D. H. Jones, P. Jennum, A. Jordanova, L. Jönsson, K. Karampampa, M. Knapp, G. Kobelt, T. Kurth, R. Lieb, M. Linde, C. Ljungcrantz, A. Maercker, B. Melin, M. Moscarelli, A. Musayev, F. Norwood, M. Preisig, M. Pugliatti, J. Rehm, L. Salvador-Carulla, B. Schlehofer, R. Simon, H.-C. Steinhausen, L. J. Stovner, J.-M. Vallat, P. V. den Bergh, J. van Os, P. Vos, W. Xu, H.-U. Wittchen, B. Jönsson, J. Olesen, Cost of disorders of the brain in Europe 2010, European Neuropsychopharmacology 21 (10) (2011) 718–779.

[3] M. E. Schwab, A. D. Buchli, Drug research: plug the real brain drain, Nature 483 (7389) (2012) 267–268.

[4] J. C. Hobart, S. J. Cano, J. P. Zajicek, A. J. Thompson, Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations, The Lancet Neurology 6 (12) (2007) 1094–1105.

[5] J. Hobart, Rating scales for neurologists, Journal of Neurology, Neurosurgery & Psychiatry 74 (suppl 4) (2003) iv22–iv26.

[6] A. Agresti, Analysis of ordinal categorical data, 2nd Edition, Wiley Series in Probability and Statistics, Hoboken, New Jersey, 2010.

[7] C. Winship, R. D. Mare, Regression models with ordinal variables, American Sociological Review 49 (4) (1984) 512–525.

[8] T. J. Hastie, J. L. Botha, C. M. Schnitzler, Regression with an ordered categorical response, Statistics in Medicine 8 (7) (1989) 785–794.

[9] S. C. Scott, M. S. Goldberg, N. E. Mayo, Statistical assessment of ordinal outcomes in comparative studies, Journal of clinical epidemiology 50 (1) (1997) 45–55.

[10] S. C. Kirshblum, W. Waring, F. Biering-Sorensen, S. P. Burns, M. Johansen, M. Schmidt-Read, W. Donovan, D. E. Graves, A. Jha, L. Jones, M. J. Mulcahey, A. Krassioukov, Reference for the 2011 revision of the international standards for neurological classification of spinal cord injury, The Journal of Spinal Cord Medicine 34 (6) (2011) 547–554.

[11] M. B. Bracken, M. J. Shepard, W. F. Collins, T. R. Holford, W. Young, D. S. Baskin, H. M. Eisenberg, E. Flamm, L. Leo-Summers, J. Maroon, L. F. Marshall, P. L. Perot, J. Piepmeier, V. K. Sonntag, F. C. Wagner, J. E. Wilberger, H. R. Winn, A randomized, Controlled Trial of Methylprednisolone or Naloxone in the Treatment of Acute Spinal-Cord Injury - Results of the Second National Acute Spinal Cord Injury Study, The New England Journal of Medicine 322 (20) (1990) 1405–1411.

[12] S. Casha, D. Zygun, M. D. McGowan, I. Bains, V. W. Yong, R. John Hurlbert, Results of a phase II placebo-controlled randomized trial of minocycline in acute spinal cord injury, Brain 135 (4) (2012) 1224–1236.

[13] C. Rudhe, H. J. A. van Hedel, Upper Extremity Function in Persons with Tetraplegia: Relationships Between Strength, Capacity, and the Spinal Cord Independence Measure, Neurorehabilitation and Neural Repair 23 (5) (2009) 413–421.

[14] D. J. Bartholomew, F. Steele, I. Moustaki, J. I. Galbraith, Analysis of multivariate social science data, 2nd Edition, Statistics in the Social and Behavioral Sciences, Chapman & Hall/CRC, London, 2008.

[15] F. Samejima, Estimation of latent ability using a response pattern of graded scores, Psychometrika Monograph 34 (4).

[16] R. D. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika 37 (1) (1972) 29–51.

[17] S. L. Sclove, Application of model-selection criteria to some problems in multivariate analysis, Psychometrika 52 (3) (1987) 333–343.

[18] T. E. Duncan, S. Dunca, An introduction to latent growth curve modeling, Behavior Therapy 35 (2) (2004) 333–363.

[19] T.-Y. Lu, W.-Y. Poon, Y.-F. Tsang, Latent growth curve modeling for longitudinal ordinal responses with applications, Computational Statistics & Data Analysis 55 (3) (2011) 1488–1497.

[20] K. A. Bollen, P. J. Curran, Latent Curve Models: a structural equation perspective, Wiley series in probability and statistics, Wiley, Hoboken, New Jersey, 2006.

[21] M. Forrest, B. Andersen, Ordinal scale and statistics in medical research., British Medical Journal 292 (6519) (1986) 537.

[22] F. H. Geisler, W. P. Coleman, G. Grieco, D. Poonian, Sygen Study Group, The Sygen® multicenter acute spinal cord injury study, Spine 26 (24S) (2001) S87–S98.

[23] D. D. Cardenas, J. Ditunno, V. Graziani, A. B. Jackson, D. Lammertse, P. Potter, M. Sipski, R. Cohen, A. R. Blight, Phase 2 trial of sustained-release fampridine in chronic spinal cord injury, Spinal Cord 45 (2) (2007) 158–168.

[24] D. P. Lammertse, L. A. T. Jones, S. B. Charlifue, S. C. Kirshblum, D. F. Apple, K. T. Ragnarsson, S. P. Falci, R. F. Heary, T. F. Choudhri, A. L. Jenkins, R. R. Betz, D. Poonian, J. P. Cuthbert, A. Jha, D. A. Snyder, N. Knoller, Autologous incubated macrophage therapy in acute, complete spinal cord injury: results of the phase 2 randomized controlled multicenter trial, Spinal Cord 50 (9) (2012) 661–671.

[25] B. Dobkin, D. Apple, H. Barbeau, M. Basso, A. Behrman, D. Deforge, J. Ditunno, G. Dudley, R. Elashoff, L. Fugate, Harkema, S., Saulino, M., Scott, M., SCILT Group, Weight-supported treadmill vs over-ground training for walking after acute incomplete SCI, Neurology 66 (4) (2006) 484–493.

[26] W. Coleman, Injury severity as primary predictor of outcome in acute spinal cord injury: retrospective results from a large multicenter clinical trial, The Spine Journal 4 (4) (2004) 373–378.

[27] L. G. Tanadini, T. Hothorn, L. A. Jones, D. P. Lammertse, R. Abel, D. Maier, R. Rupp, N. Weidner, A. Curt, J. D. Steeves, Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders Prediction-Based Stratification of Participants With Incomplete Cervical Spinal Cord Injury, Neurorehabilitation and Neural Repair 29 (9) (2015) 867–877.

[28] J. D. Steeves, J. K. Kramer, J. W. Fawcett, J. Cragg, D. P. Lammertse, A. R. Blight, R. J. Marino, J. F. Ditunno, W. P. Coleman, F. H. Geisler, Extent of spontaneous motor recovery after traumatic cervical sensorimotor complete spinal cord injury, Spinal Cord 49 (2) (2011) 257–265.

[29] P. M. W. Bath, L. J. Gray, T. Collier, S. Pocock, j. Carpenter, Can We Improve the Statistical Analysis of Stroke Trials? Statistical Reanalysis of Functional Outcomes in Stroke Trials, Stroke 38 (6) (2007) 1911–1915.

[30] A. I. Maas, G. D. Murray, B. Roozenbeek, H. F. Lingsma, I. Butcher, G. S. McHugh, J. Weir, J. Lu, E. W. Steyerberg, IMPACT Study Group, Advancing care for traumatic brain injury: findings from the IMPACT studies and perspectives on future research, The Lancet Neurology 12 (12) (2013) 1200–1210.

[31] G. Rasch, Probabilistic Models for Some Intelligence and Attainment Tests, Vol. Paedagogiske Institut, Copenhagen, 1960.

[32] R. D. McKelvey, W. Zavoina, A statistical model for the analysis of ordinal level dependent variables, Journal of Mathematical Sociology 4 (1975) 103–120.

[33] P. McCullagh, Regression models for Ordinal Data, Journal of the Royal Statistical Society 42 (2) (1980) 109–142.